

## WEST Search History

DATE: Tuesday, June 21, 2005

Hide?	Set Name	Query	Hit Count
		<i>DB=PGPB,USPT,USOC,EPAB,JPAB,DWPI,TDBD; PLUR=YES; OP=AND</i>	
<input type="checkbox"/>	L1	(johnson or adamou).in. and strep\$	637
<input type="checkbox"/>	L2	L1 and pneumoni\$	153
<input type="checkbox"/>	L3	L2 and (\$triad or coil\$)	15

END OF SEARCH HISTORY

20040072160. 22 May 02. 15 Apr 04. Molecular toxicology modeling. Mendrick, Donna, et al. 435/6; 435/91.2 436/84 C12Q001/68 C12P019/34 G01N033/20.

---

❑ 2. [20040052781](#). 14 Apr 03. 18 Mar 04. Vaccine compositions comprising Streptococcus pneumoniae polypeptides having selected structural motifs. Johnson, Leslie S., et al. 424/130.1; 424/185.1 435/100 A61K039/395 A61K039/00.

---

❑ 3. [20040005331](#). 13 Mar 03. 08 Jan 04. Vaccine compositions comprising Streptococcus pneumoniae polypeptides having selected structural motifs. Johnson, Leslie S., et al. 424/190.1; 530/350 536/23.7 A61K039/02 C07H021/04 C07K014/315.

---

❑ 4. [20040001836](#). 14 Apr 03. 01 Jan 04. Vaccine compositions comprising streptococcus pneumoniae polypeptides having selected structural motifs. Johnson, Leslie S., et al. 424/165.1; 424/190.1 A61K039/40 A61K039/02.

---

❑ 5. [20030138447](#). 25 Sep 02. 24 Jul 03. Derivatives of choline binding proteins for vaccines. Wizemann, Theresa M., et al. 424/190.1; 424/234.1 A61K039/02.

---

❑ 6. [20030125337](#). 26 Mar 02. 03 Jul 03. Inhibitors of multidrug transporters. Markham, Penelope N., et al. 514/253.08; 514/312 514/410 514/415 A61K031/496 A61K031/4709 A61K031/47 A61K031/405.

---

❑ 7. [6863893](#). 25 Sep 02; 08 Mar 05. Derivatives of choline binding proteins for vaccines. Wizemann, Theresa M., et al. 424/190.1; 424/184.1 424/234.1 424/237.1 424/244.1 514/2 514/8 530/300 530/350. A61K03902 A61K039085 A16K03800 A16K03816 C07K01400.

---

❑ 8. [6833356](#). 25 Aug 00; 21 Dec 04. Pneumococcal protein homologs and fragments for vaccines. Koenig, Scott, et al. 514/12; 424/130.1 424/184.1 424/243.1 424/244.1 514/2 530/350 536/23.1. C07K014/00 A61K038/16.

---

❑ 9. [6582706](#). 21 Dec 99; 24 Jun 03. Vaccine compositions comprising Streptococcus pneumoniae polypeptides having selected structural MOTIFS. Johnson, Leslie S., et al. 424/244.1; 424/184.1 424/185.1 424/190.1 424/237.1 435/320.1 435/69.1 530/350 536/23.1 536/23.7. A61K039/09.

---

❑ 10. [6503511](#). 06 Apr 99; 07 Jan 03. Derivatives of choline binding proteins for vaccines. Wizemann, Theresa M., et al. 424/190.1; 424/184.1 424/234.1 424/237.1 424/244.1 514/2 514/8 530/300 530/350. A61K039/02 A61K039/085 A61K038/00 A61K038/16 C07K014/00.

---

❑ 11. [6362229](#). 17 Aug 00; 26 Mar 02. Inhibitors of multidrug transporters. Markham, Penelope N., et al. 514/596; 514/311 546/112 546/134 546/139 546/152 564/48 564/53. A61K031/17 C07C275/06 C07C275/30.

---

❑ 12. [6326391](#). 02 Dec 99; 04 Dec 01. Inhibitors of multidrug transporters. Markham, Penelope N., et al. 514/410; 514/412 548/420 548/452 548/469. A61K031/404 A61K031/402 A61K031/403 C07D209/04.

---

❑ 13. [WO 200037105A](#). Vaccine useful for prophylaxis and treatment of pneumococcal infections such as otitis media, nasopharyngeal and bronchial infections, comprises Streptococcus pneumoniae proteins. ADAMO, J E, et al. A61K038/00 A61K039/00 A61K039/02 A61K039/09 A61K039/395 A61K039/40 A61P031/04 A61P031/10 C07H021/04 C07K014/315 C07K014/315.

---

☐ 14. 3207750. 21 Sep 65. Derivatives of decoyinine. DE BOER CLARENCE; JOHNSON LE ROY E ; EBLE THOMAS E ; HERMAN HOEKSEMA. 536/27.5; 435/88 435/898.

---

☐ 15. 3094460. 18 Jun 63. Decoyinine. DE BOER CLARENCE; JOHNSON LE ROY E ; EBLE THOMAS E ; HERMAN HOEKSEMA. 424/118; 424/114 435/119 435/88 435/898 514/29 514/31 514/45 536/27.23 536/27.5.

---

[Generate Collection](#)[Print](#)

Terms	Documents
L2 and (\$triad or coil\$)	15

[Prev Page](#)[Next Page](#)[Go to Doc#](#)



US006833356B1

(12) **United States Patent**  
Koenig et al.

(10) Patent No.: **US 6,833,356 B1**  
(45) Date of Patent: **Dec. 21, 2004**

(54) **PNEUMOCOCCAL PROTEIN HOMOLOGS  
AND FRAGMENTS FOR VACCINES**

(75) Inventors: **Scott Koenig**, Rockville, MD (US); **Jon Heinrichs**, North Potomac, MD (US); **Leslie S. Johnson**, Germantown, MD (US); **John E. Adamou**, Germantown, MD (US)

(73) Assignee: **Medimmune, Inc.**, Gaithersburg, MD (US)

(\*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 197 days.

(21) Appl. No.: **09/645,835**

(22) Filed: **Aug. 25, 2000**

#### Related U.S. Application Data

(60) Provisional application No. 60/150,750, filed on Aug. 25, 1999.

(51) Int. Cl.<sup>7</sup> ..... **C07K 14/00; A61K 38/16**

(52) U.S. Cl. .... **514/12; 514/2; 530/350; 424/184.1; 424/130.1; 424/243.1; 424/244.1; 536/23.1**

(58) Field of Search ..... **514/12, 2; 530/350, 530/23.1; 424/184.1, 130.1, 243.1, 244.1, 185.1; 536/23.1**

(56) **References Cited**

#### U.S. PATENT DOCUMENTS

4,694,073 A \* 9/1987 Bentle et al. .... 530/399

2003/0031682 A1 \* 2/2003 Brodeur et al. .... 424/190.1

#### FOREIGN PATENT DOCUMENTS

WO	WO 98/18930	5/1998
WO	WO 99/42588	8/1999
WO	WO 00/06736	2/2000

#### OTHER PUBLICATIONS

Spellerberg et al., Lmb, a protein with similarities to the Lral adhesin family, mediates attachment of streptococcus agalactiae to human laminin. *Infection and Immunity* Feb. 1999, vol. 67 871-878.\*

\* cited by examiner

Primary Examiner—Robert A. Wax

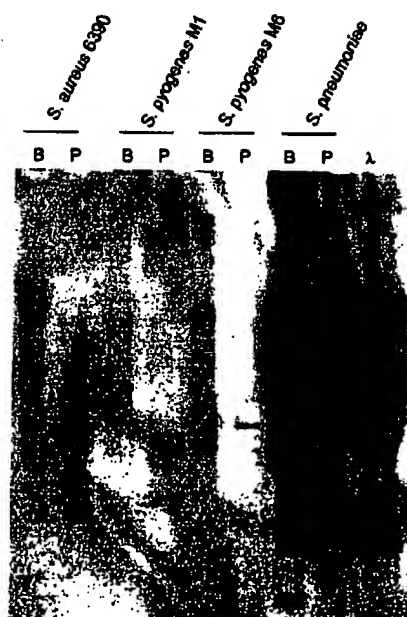
Assistant Examiner—Chih-Min Kam

(74) Attorney, Agent, or Firm—Elliott M. Olstein; Alan J. Grant

(57) **ABSTRACT**

The invention is directed to isolated polypeptides bearing sequence homology to the Sp36 protein found in pneumococcal organisms, such as *Streptococcus pneumoniae*. Polynucleotides encoding such polypeptides are also disclosed. The invention also relates to antibodies specific for the disclosed polypeptides and to uses of such antibodies in the treatment of diseases caused by staphylococci as well as group A and B streptococci. In addition, the invention relates to the use of the disclosed polypeptides in compositions and as vaccines and for prophylactic uses such as in vaccination of animals, especially humans, against a wide variety of streptococcal, staphylococcal and other diseases.

**8 Claims, 9 Drawing Sheets**



-continued

---

Val	Glu	His	Pro	Asp	Glu	Arg	Pro	His	Ser	Asn	Asp	Gly	Trp	Gly	Asn	
690										695					700	
Ala	Ser	Glu	His	Val	Leu	Gly	Lys	Lys	Asp	His	Ser	Glu	Asp	Pro	Asn	
705					710					715					720	
Lys	Asn	Phe	Lys	Ala	Asp	Glu	Glu	Pro	Val	Glu	Glu	Thr	Pro	Ala	Glu	
				725					730					735		
Pro	Glu	Val	Pro	Gln	Val	Glu	Thr	Glu	Lys	Val	Glu	Ala	Gln	Leu	Lys	
				740				745					750			
Glu	Ala	Glu	Val	Leu	Leu	Ala	Lys	Val	Thr	Asp	Ser	Ser	Leu	Lys	Ala	
		755					760					765				
Asn	Ala	Thr	Glu	Thr	Leu	Ala	Gly	Leu	Arg	Asn	Asn	Leu	Thr	Leu	Gln	
		770				775						780				
Ile	Met	Asp	Asn	Asn	Ser	Ile	Met	Ala	Glu	Ala	Glu	Lys	Leu	Leu	Ala	
785					790					795					800	
Leu	Leu	Lys	Gly	Ser	Asn	Pro	Ser	Ser	Val	Ser	Lys	Glu	Lys	Ile	Asn	
				805					810					815		

---

What is claimed is:

1. An isolated polypeptide comprising an amino acid sequence with at least 95% sequence identity to the sequence of SEQ ID NO: 4 and wherein said polypeptide binds to an antibody that is specific for Sp36 (SEQ ID NO: 7).

2. An isolated polypeptide comprising an amino acid sequence with at least 95% sequence identity to a sequence selected from the group consisting of SEQ ID NO: 2 and 4 wherein said polypeptide is identical to that found in an organism selected from the group consisting of Group A streptococci and *Staphylococcus aureus* and wherein said polypeptide binds to an antibody that is specific for Sp36 (SEQ ID NO: 7).

3. The isolated polypeptide of claim 2 wherein said Group A organism is *Streptococcus pyogenes*.

4. The isolated polypeptide of claim 2 wherein said organism is *Staphylococcus aureus*.

25 5. An isolated polypeptide comprising an amino acid sequence at least 95% identical to the sequence of SEQ ID NO: 4 and wherein said polypeptide has a sequence with at least 12.6% sequence identity to the amino acid sequence of the Sp36 protein (SEQ ID NO: 7) of *Streptococcus pneumoniae* and wherein said isolated polypeptide binds to an antibody that is specific for Sp36.

30 6. An isolated polypeptide comprising the sequence of SEQ ID NO: 2 wherein said isolated polypeptide binds to an antibody that is specific for Sp36 (SEQ ID NO: 7) of *Streptococcus pneumoniae*.

35 7. An isolated polypeptide comprising the amino acid sequence of SEQ ID NO: 2.

40 8. An isolated polypeptide comprising the amino acid sequence of SEQ ID NO: 4.

\* \* \* \* \*



Get  for  Go ? Site search


[EBI Home](#)
[About EBI](#)
[Groups](#)
[Services](#)
[Toolbox](#)
[Databases](#)
[Downloads](#)

DATABASE BROWSING

## EBI Dbfetch

```

ID  AF291695    standard; genomic DNA; PRO; 2541 BP.
XX
AC  AF291695;
XX
SV  AF291695.1
XX
DT  19-MAR-2001 (Rel. 67, Created)
DT  19-MAR-2001 (Rel. 67, Last updated, Version 1)
XX
DE  Streptococcus pneumoniae pneumococcal histidine triad A protein (phtA)
DE  gene, complete cds.
XX
KW  .
XX
OS  Streptococcus pneumoniae
OC  Bacteria; Firmicutes; Lactobacillales; Streptococcaceae; Streptococcus.
XX
RN  [1]
RP  1-2541
RX  DOI; 10.1128/IAI.69.3.1593-1598.2001
RX  PUBMED; 11179332.
RA  Wizemann T.M., Heinrichs J.H., Adamou J.E., Erwin A.L., Kunsch C.,
RA  Choi G.H., Barash S.C., Rosen C.A., Masure H.R., Tuomanen E., Gayle A.,
RA  Brewah Y.A., Walsh W., Barren P., Lathigra R., Hanson M., Langermann S.,
RA  Johnson S., Koenig S.;
RT  "Use of a whole genome approach to identify vaccine molecules affording
RT  protection against Streptococcus pneumoniae infection.";
RL  Infect. Immun. 69(3):1593-1598 (2001).
XX
RN  [2]
RP  1-2541
RA  Choi G.H.;
RT  ;
RL  Submitted (01-AUG-2000) to the EMBL/GenBank/DDBJ databases.
RL  Molecular Biology, Human Genome Sciences, Inc., 9410 Key West Ave.,
RL  Rockville, MD 20850, USA
XX
FH  Key          Location/Qualifiers
FH
FT  source       1..2541
FT                /db_xref="taxon:1313"
FT                /mol_type="genomic DNA"
FT                /organism="Streptococcus pneumoniae"
FT                /strain="N4"
FT  CDS         91..2541
FT                /codon_start=1
FT                /db_xref="InterPro:IPR006270"
FT                /db_xref="UniProt/TrEMBL:Q9AHT9"
FT                /note="PhtA"
FT                /transl_table=11
FT                /gene="phtA"
FT                /product="pneumococcal histidine triad A protein"
FT                /protein_id="AAK19155.1"

```

```

FT      /translation="MKINKKYLVGSAALILSVCSYELGLYQARTVKENNRVSYIDGKQ
FT      ATQKTENLTPDEVSKREGINAEQIVIKITDQGYVTSHGDHYHYNGKVPYDAI ISEELL
FT      MKDPNYKLKDEDIVNEVKGGYVIKVDGKYVYLKDAAHADNVRTKEE INRQKQEHSHR
FT      EGGTPRNDGAVALARSQGRYTTDDGYIFNASDI IEDTGDAYIVPHGDHYHYIPKNELSA
FT      SELAAAEAFLSGRGNLSNSRTYRRQNSDNTSRTNWVPSVSNPGTTNTNTSNNSNTNSQA
FT      SQSNDIDSLLKQLYKLPLSQRHVESDGLVFDPAQITTSRTARGVAVPHGDHYHFIPYSQM
FT      SELEERIARI IPLRYRSNHWPDSRPEQSPQPTPEPSPGPQPAPNLKIDSNSSLVSQ
FT      VRKVGEGYVFEEKGISRYVFAKDLPSSETVKNLESKLSKQESVSHTLTAKKENVAPRDQE
FT      FYDKAYNLLTEAHKALFXNKGKNSDFQALDKLLERLNDESTNKEKLVDDLLAFLAPITH
FT      PERLGKPNQIEYTEDEVRIAQLADKYTTSDGYIFDEHDI ISDEGDAYVTPHMGHSHWI
FT      GKDSLSDKEKVAAQAYTKEKGILPPSPDADVKNPTGDSAAAI YNRVKGEKRI PLVRLP
FT      YMVEHTVEVKNGNLI I PHKDHYHNI KFAWFDDHTYKAPNGYTLEDLFATIKYYVEHPDE
FT      RPHSNDGWGNASEHVLGKKDHSEDPNKNFKADEEPVEETPAEPEVPQVETEKVEAQLKE
FT      AEVLLAKVTDSSSLKANATETLAGLRNLTQIMDNNS IMAEAEKLLALLKGSNPSSVSK
FT      EKIN"
XX
SQ

```

```



Sequence 2541 BP; 888 A; 476 C; 516 G; 660 T; 1 other;
taaactatta accagttaag taatagagag gagtttctgc aatttagaaa tgaattgcaa      60
ctagaaatat caaatagaaa gagagtttcg atgaaaatta ataagaaata cttgtgtggt      120
tctgcgcgag ctttgatttt aagtgtttgt tcttacgagt tgggactgta tcaagctaga      180
acggttaagc aaaaataatcg tgtttcctat atagatggaa aacaagcgac gcaaaaaacg      240
gagaatttga ctcctgatga ggtagcaag cgtgaaggaa tcaatgctga gcaaatcgct      300
atcaagataa cagaccaagg ctatgtcact tcacatggcg accactatca ttattacaat      360
ggtaagggttc cttatgacgc tatcatcagt gaagaattac tcatgaaaga tccaaactat      420
aagctaaaag atgaggatat tgttaatgag gtcaaggggt gatatgttat caaggtagat      480
ggaaaatact atgtttacct taaggatgct gccacgcgag ataacgtccg tacaaaagag      540
gaaatcaatc gacaaaaaca agagcatagt caacatcggt aaggtggaac tccaagaaac      600
gatggtgctg ttgccttggc acgttcgcaa ggacgctata ctacagatga tggttatatc      660
tttaattgctt ctgatatcat agaggatact ggtgatgctt atatcgttcc tcatggagat      720
cattaccatt acattcctaa gaatgagtta tcagctagcg agttggctgc tgcagaagcc      780
ttcctatctg gtcgaggaaa tctgtcaaat tcaagaacct atcgccgaca aaatagcgat      840
aacacttcaa gaacaaactg ggtaccttct gtaagcaatc caggaaactac aaatactaac      900
acaagcaaca acagcaacac taacagtcaa gcaagtcaaa gtaatgacat tgatagtctc      960
ttgaaacagc tctacaaact gcctttgagt caacgacatg tagaatctga tggccttgct      1020
tttgatccag cacaaatcac aagtcgaaca gctagagggt ttgcagtgcc acacggagat      1080
cattaccact tcatccctta ctctcaaatg tctgaattgg aagaacgaaat cgctcgatt      1140
attccccttc gttatcggtc aaaccattgg gtaccagatt caaggccaga acaaccaagt      1200
ccacaaccga ctccggaacc tagtccaggc ccgcaacctg caccaaatct taaaatagac      1260
tcaaattcct ctttggttag tcagctggta cgaaaagtgg gggaaggata tgtattcgaa      1320
gaaaagggca tctctcgtaa tgtctttgcg aaagatttac catctgaaac tgttaaaaaat      1380
cttgaaagca agttatcaaa acaagagagt gtttcacaca ctttaactgc taaaaagaa      1440
aatgttgctc ctggtgacca agaattttat gataaagcat ataactgtgt aactgaggct      1500
cataaagcct tgtttgnaaa taagggtcgt aattctgatt tccaagcctt agacaaatta      1560
ttagaacgct tgaatgatga atcgactaat aaagaaaaat tggtagatga tttattggca      1620
ttcctagcac caattaccac tccagagcga cttggcaaac caaattctca aattgagtat      1680
actgaagacg aagttcgtat tgctcaatta gctgataagt atacaacgtc agatggttac      1740
atttttgatg aacatgatat aatcagtgat gaaggagatg catatgtaac gcctcatatg      1800
ggccatagtc actggattgg aaaagatagc ctttctgata aggaaaaagt tgcagctcaa      1860
gcctatacta aagaaaaagg tatcctacct ccatctccag acgcagatgt taaagcaaat      1920
ccaactggag atagtgcagc agctatttac aatcgtgtga aaggggaaaa acgaattcca      1980
ctcgttcgac ttccatatat ggttgagcat acagttgagg ttaaaaaacg taatttgatt      2040
attcctcata aggatcata ccataatatt aaatttgctt ggtttgatga tcacacatac      2100
aaagctccaa atggctatac cttggaagat ttgtttgcga cgattaagta ctacgtagaa      2160
caccctgacg aacgtccaca ttctaagatg ggatggggca atgccagtga gcatgtgtta      2220
ggcaagaaaag accacagtga agatccaaat aagaacttca aagcggatga agagccagta      2280
gaggaaacac ctgctgagcc agaagtcctt caagtagaga ctgaaaaagt agaagcccaa      2340
ctcaaagaag cagaagtttt gcttgcgaaa gtaacggatt ctagtctgaa agccaatgca      2400
acagaaactc tagctggttt acgaaataat ttgactcttc aaattatgga taacaatagt      2460
atcatggcag aagcagaaaa attacttgcg ttgttaaaag gaagtaatcc ttcattctgta      2520
agtaaggaaa aaataaacta a

```

//

[0077] The identification of multiple coil structures of alpha helical amino acid segments in the S. pneumoniae polypeptides according to the invention may be determined by the location of proline rich areas with respect to one another. Further the "X" area optionally located between two or more alpha-helical structures can play a part in the formation of a coil within a coil structure. Standard three-dimensional protein modeling may be utilized for determining the relative shape of such structures. An example of a computer program, the Paircoil Scoring Form Program ("PairCoil program"), useful for such three-dimensional protein modelling is described by Berger et al. in the Proc. Natl. Acad. of Sci. (USA), 92:8259-8263 (August 1995). The PairCoil program is described as a computer program that utilizes a mathematical algorithm to predict locations of coiled-coil regions in amino acid sequences. A further example of such a computer program is described by Wolf et al., Protein Science 6:1179-1189 (June 1997) which is called the Multicoil Scoring Form Program ("Multicoil program"). The MultiCoil program is based on the PairCoil algorithm and is useful for locating dimeric and trimeric coiled coils.



Search Results - Record(s) 1 through 5 of 5 returned.

- ☐ 1. 6773880. 03 Jan 01; 10 Aug 04. Streptococcus pneumoniae 37-kDa surface adhesion A protein. Sampson; Jacquelyn, et al. 435/6; 536/23.7 536/24.32 536/24.33. C12Q001/68.
- ☐ 2. 6582706. 21 Dec 99; 24 Jun 03. Vaccine compositions comprising Streptococcus pneumoniae polypeptides having selected structural MOTIFS. Johnson; Leslie S., et al. 424/244.1; 424/184.1 424/185.1 424/190.1 424/237.1 435/320.1 435/69.1 530/350 536/23.1 536/23.7. A61K039/09.
- ☐ 3. 6406883. 25 Sep 98; 18 Jun 02. Lmb gene of Streptococcus agalactiae. Luticken; Rudolf, et al. 435/69.1; 424/244.1 435/243 435/252.3 435/253.4 435/320.1 435/69.3 536/23.7. C12P021/06.
- ☐ 4. 6217884. 28 Dec 98; 17 Apr 01. Streptococcus pneumoniae 37-kDa surface adhesin a protein. Sampson; Jacquelyn S., et al. 424/244.1; 424/184.1 424/190.1 424/200.1 435/69.1 435/69.3 435/71.1 530/350 536/23.7. A61K039/09.
- ☐ 5. 5854416. 17 Sep 96; 29 Dec 98. Streptococcus pneumoniae 37-KDA surface adhesin a protein and nucleic acids coding therefor. Sampson; Jacquelyn S., et al. 536/23.7; 424/244.1 435/320.1 536/23.1. C07H021/04.

Terms	Documents
lxxc	5

[Prev Page](#) [Next Page](#) [Go to Doc#](#)

tr Q97QM9 Conserved domain protein [SP1174] [Streptococcus pneumoniae] 819 AA

align

Score = 423 bits (1087), Expect = e-116

Identities = 211/357 (59%), Positives = 271/357 (75%), Gaps = 11/357 (3%)

```
Query: 1  MKFSKKYIAAGSAVIVSLSLCAYALNQHRS-QENKDNRRVSYVDGSQSSQKSENLTDPQV 59
          MK +KKY+A GS  +++LS+C+Y L +++++ Q+ K++NRV+Y+DG Q+ QK+ENLTPD+V
Sbjct: 1  MKINKKYLA-GSVAVLALSVCSYELGRYQAGQDKKESNRVAYIDGDQAGQKAENLTPDEV 59

Query: 60  SQKEGIQAEQIVIKITDQGYVTSHGDHYHYNGKVPYDALFSEELLMKDPNYQLKDADIV 119
          S++EGI AEQIVIKITDQGYVTSHGDHYHYNGKVPYDA+ SEELLMKDPNYQLKD+DIV
Sbjct: 60  SKREGINAEQIVIKITDQGYVTSHGDHYHYNGKVPYDAIIEELLMKDPNYQLKDS DIV 119

Query: 120  NEVKGGYIIKVDGKYYVYLKDAAHADNVRTKDEINRQKQEHVKD-NEKVNSNVAVARSQG 178
          NE+KGGY+IKV+GKYYVYLKDAAHADN+RTK+EI RQKQE  + N + ++ VA AR+QG
Sbjct: 120  NEIKGGYVIKVGNGKYYVYLKDAAHADNIRTKEEIKRQKQERSHNHNSRADNAVAARAAG 179

Query: 179  RYTTNDGYVFNPADIIEDTGDAYIVPHGGHYHYIPXXXXXXXXXXXXXXXXXXXXNMQPSQ 238
          RYTT+DGY+FN +DIIEDTG+AYIVPHG HYHYIP                      Q S+
Sbjct: 180  RYTDDGYIFNASDIIEDTGDAYIVPHGDHYHYIP--KNELSASELAAAEAYWNGKQGSR 237

Query: 239  LSYSSSTASDNNTQ---SVAKGSTSKPA---NKSENLSLLKELYDSPAQRYSES DGLVF 292
          S SS+ + N Q  S      T P      N+ EN+ SLL+ELY P ++R+ ESDGL+F
Sbjct: 238  PSSSSSYNANPAQPRLSHNHNLTVPTTYHQNGENISSLLRELYAKPLSERHVESDGLIF 297

Query: 293  DPAKIISRTPNGVAIPHGDHYHFIPYSKLSALEEKIARRVPISGTGSTVSTNAKPNE 349
          DPA+I SRT  GVA+PHG+HYHFIPY ++S LE++IAR +P+  +  +++P E
Sbjct: 298  DPAQITSRTARGVAVPHGNHYHFIPYEQMSELEKRIARIIPLYRSNHWVPDSRPEE 354
```

Submission	Matches on query sequence		Mat
	1	500	
Q8DQ87			
Q6MNP7			
Q9ANY1			
Q6MNP5			
Q8CNR4			
Q8DPQ2			
Q9AG74			
Q9AHT9			
Q8DQ88			
Q6T8D7			
Q97QH8			
Q9ANY2			
Q97QM9			
Q9ANY3			
Q6MNP3			
Q6MNP8			
Q6MNP1			
Q6MNP5			
Q6MNP9			
Q6T384			
Q6MNP8			
Q6MNP6			
Q8NZ82			
Q8E4U1			
Q8DZ81			
Q9ZHC7			
Q99XV4			
Q8K5Q1			
Q93GT5			
Q8E338			
Q877Y2			
Q9AE21			
Q8DQ86			
Q8E829			
Q8E5R2			
Q8P8G5			
Q8K714			
Q79XH7			
Q99Z76			
Q6HCJ8			
IGA4_HAEIN			
Q8MNS8			
Q8ISF7			
Q8ISF6			
NFM_CHICK			
Q97QP7			
Q8IB63			
Q869E1			
Q9VC88			
Q839C3			
Q73793			
Q9FN97			
Q963T1			
Q87594			
Q6PK21			
OGFR_HUMAN-2			
OGFR_HUMAN			
Q6HBX5			
Q7RQS8			
Q77328			
Q9L4Z1			
Q9VGN4			
Q98387			
Q898B8			
Q6R4Z8			
Q8MNP1			
Q8IBL1			
QDP2_STREP			
Q33741			
Q8I1H5			
Q87593			
TRDN_RABIT			
Q7SKH9			
Q8DPR5			
Q59947			
Q28688			
Q94674			
Q6BLN8			
TRDN_RABIT-4			
TRDN_RABIT-6			
Q54875			
Q9GUY4			
Q9GTX2			
Q8IJ56			
Q6FNC8			
Q6CTI8			

National Application No.

PCT/CA 99/01218

## CLASSIFICATION OF SUBJECT MATTER

IPC 7 C12N15/31 C12N15/62 C07K14/315 A61K39/09

According to International Patent Classification (IPC) or to both national classification and IPC

## B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

IPC 7 C12N C07K A61K

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)

WPI Data, PAJ, CAB Data, STRAND

## C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
------------	--	-----------------------

X	WO 98 18930 A (HUMAN GENOME SCIENCES INC ;CHOI GIL H (US); HROMOCKYJ ALEX (US); J) 7 May 1998 (1998-05-07) cited in the application SP103; SEQ ID NOs. 181 and 182; page 85, line 14 - line 42; claims 1-21; table I SEQ ID Nos. 65 and 66;	1-12
---	---	------

X	WO 98 18931 A (DOUGHERTY BRIAN A ;HUMAN GENOME SCIENCES INC (US); ROSEN CRAIG A ( ) 7 May 1998 (1998-05-07) SEQ ID No. 192 claims 1-20 SEQ ID No. 94	1-12
---	--	------

-/-

☒ Further documents are listed in the continuation of box C.☒ Patent family members are listed in annex.

## \* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" documents which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"Z" document member of the same patent family

Date of the actual completion of the international search

28 June 2000

Date of mailing of the international search report

24.07.00

Name and mailing address of the ISA

European Patent Office, P.B. 5518 Patentlaan 2  
NL - 2280 HV Rijswijk  
Tel. (+31-70) 340-2040, Tx. 31 651 epo nl,  
Fax: (+31-70) 340-3018

Authorized officer

Hornig, H

Table 1

62

AGAAGCCTATTGGAATGGGAAGCAGGGATCTCGTCCTTCTTCAAGTTCTAGTTATAATGCAAAATCCAGC  
TCAACCAAGATTGTGAGAGAACCACAATCTGACTGTCACTCCAATTATCATCAAAATCAAGGGGAAAA  
CATTTCAAGCCTTTTACGTGAATTGTATGCTAAACCCCTTATCAGAACGCCATGTGGAATCTGATGGCCT  
TATTTTCGACCCAGCGCAAATCACAAGTCGAACCGCCAGAGGTGTAGCTGTCCCTCATGGTAACCATTA  
CCACTTTATCCCTTATGAACAAATGTCTGAATTGGAAAAACGAATTGCTCGTATTATCCCTTCGTTA  
TCGTTCAAACCATTTGGGTACCAGATTCAAGACCAGAACCAAGTCCACAATCGACTCCGGAACTTAG  
TCCAAGTCCGCAACCTGCACCAAATCCTCAACCAGCTCCAAGCAATCCAATTGATGAGAAATTGGTCAA  
AGAAGCTGTTTGAAAGTAGGCGATGGTTATGTCTTTGAGGAGAATGGAGTTTCTCGTTATATCCCAGC  
CAAGGATCTTTTACGAGAAACAGCAGCAGGCATTGATAGCAAACTGGCCAAGCAGGAAAGTTTATCTCA  
TAAGCTAGGAGCTAAGAAAACGACCTCCCATCTAGTGATCGAGAATTTTACAATAAGGCTTATGACTT  
ACTAGCAAGAATTCACCAAGATTTACTTGATAATAAAGGTCGACAAGTTGATTTTGAGGCTTTGGATAA  
CCTGTTGGAACGACTCAAGGATGTCNCAAGTGATAAAGTCAAGTTAGTGGANGATATTCTTGCCTTCTT  
AGCTCCGATTTCGTATCCAGAACGTTTAGGAAAACCAAATGCGCAAAATTACCTACACTGATGATGAGAT  
TCAAAGTAGCCAAGTTGGCAGGCAAGTACACAACAGAAGACGGTTATATCTTTGATCCTCGTGATATAAC  
CAGTGATGAGGGGGATGCCTATGTAACCTCCACATATGACCCATAGCCACTGGATTAAAAAGATAGTTT  
GTCTGAAGCTGAGAGAGCGGCAGCCAGGCTTATGCTAAAGAGAAAGGTTTGACCCCTCCTTCGACAGA  
CCATCAGGATTACAGAAATACTGAGGCAAAAGGAGCAGAAGCTATCTACAACCGCGTGAAGCAGCTAA  
GAAGGTGCCACTTGATCGTATGCCTTACAATCTTCAATATACTGTAGAAGTCAAAAACGGTAGTTTAAT  
CATACCTCATTATGACCATTAACATAACATCAAATTTGAGTGGTTTGACGAAGGCCTTTATGAGGCACC  
TAAGGGGTATACTCTTGAGGATCTTTTGGCGACTGTCTCAAGTACTATGTGCAACATCCAAACGAACGTC  
GCATTGAGATAATGGTTTGGTAAACGCTAGCGACCATGTTCAAAGAAAACAAAATGGTCAAGCTGATAC  
CAATCAAACGGAAAAACCAAGCGAGGAGAAACCTCAGACAGAAAACCTGAGGAAGAAACCCCTCGAGA  
AGAGAAACCGCAAGCGAGAGAAACAGAGTCTTCAAAACCAACAGAGGAACAGAAAGATCACCAGAGGA  
ATCAGAAGAACCTCAGGTCGAGACTGAAAAGGTTGAAGAAAACCTGAGAGAGGCTGAAGATTTACTTGG  
AAAAATCCAGGAT

## SP042 amino acid (SEQ ID NO:66)

CSYELGRHQAGQVKESNRVSYIDGDQAGQKAENLTPDEVSKREGINAEQXVIKITDQGYVTSBGDHYH  
YNGKVPYDAIISEELMKDPNYQLKDSDIVNEIKGGYVIKVNKYVYLKDAHADNIRTKEEIKRQK  
QERSHNHNSRADNAVAARAQGRYTTDDGYIFNASDIIEDTGDAYIVPHGDHYHIYPKNLSASELAAA  
EAYWNGKQSGSRPSSSSSYNANPAQPRLESENHNLTVPTTHQNGENISSLLRELYAKPLSERHVESDGL  
IFDPAQITSRTARGVAVPHGNHYHFIPEEQMSELEKRIARIIPLYRNSNHWPDSRPEQPSQSTPEPS  
PSPQPAPNPQPAPSNPIDEKLVKEAVRKVGQGVFEENGVSRYIPAKDLAETAAGIDSKLAKQESLSH  
KLGAKKTDLPSSDREFYNKAYDLLARIHQDLLDNKGRQVDFEALDNLLERLKDVSXKVLVXDILAF  
APIRHPERLGLKPNQITTYTDEIQVAKLAGKYTTEDGYIFDPRDITSDGEDAYVTPHMTSHWIKKDSL  
SEAERAAQAYAKEKGLTPSTDHQSNGTEAKGAEAIYNRVKAAKKVPLDRMPYNLQYTVVKNGLSLI  
IPHYDHYHNKIFEFWDEGLYEAPKGYTLEDLLATVKYVVEHPNERPHSDNGFGNASDHVQRNKGQADT  
NQTEKPSSEKPKTEKPEEETPREKPKQSEKPEPKPTEPEEPESESEEPQVETEKVEEKLREAEDLLG  
KIQD

## SP043 nucleotide (SEQ ID NO:67)

TTATAAGGGTGAATTAGAAAAAGGATACCAATTTGATGGTTGGGAAATTTCTGGTTTCAAGGTAAAAA  
AGACGCTGGCTATGTTATTAATCTATCAAAAGATACCTTTATAAAACCTGTATTCAGAAAAATAGAGGA  
GAAAAAGGAGGAAGAAAAATAAACCTACTTTTGATGTATCGAAAAAGAAAGATAACCCACAAGTAAACCA  
TAGTCAATTAATGAAAGTCACAGAAAAGAGGATTTACAAAGAGAAGAGCATTCACAAAAATCTGATTC  
AACTAAGGATGTACAGCTACAGTTCTTGATAAAAAACAATATCAGTAGTAAATCAACTACTAACAATCC  
TAATAAG

## SP043 amino acid (SEQ ID NO:68)

YKGELEKGYQFDGWEISGFEGKKDAGYVNLKDTFIKPVFKKIEEKKEENKPTFDVSKKKDNFPQVNH  
SQLNESHKEDLQREEHSQKSDSTKDVATVLDKNNISSKSTNNPNK

## SP044 nucleotide (SEQ ID NO:69)

GAATGTTTCAGGCTCAAGAAAGTTCAGGAAATAAAATCCACTTTATCAATGTTCAAGAAAGGTGGCAGTGA  
TGCGATTATCTTGAAGCAATGGACATTTTGCCATGGTGGATACAGGAGAAGATTATGATTTCACAGA  
TGGAAAGTGATTCTCGCTATCCATGGAGAGAAGGAATTGAAACGCTTTATAAGCATGTTCTAACAGACCG  
TGTCTTTTCGTCGTTTGAAGGAATTGGGTGTCAAAAACCTTGATTTTATTTTGGTGACCCATACCCACAG  
TGATCATATTGGAAATGTTGATGAATTACTGTCTACCTATCCAGTTGACCGAGTCTATCTTAAGAAATA

1157

TGTCAGAATT AACATCTCCA AACGCTGTTT TTGAATCGGT CATTCTGATA CCATTTTCTG 10200  
CACAATAAAC CAATACACGA TTATAGGCTT CTGTAGATTT AACCACTATA TACAATTCAA 10260  
TCATTTTAGA ACGATTTTGC AGATATTTT TTAGTGTTG GAACATGGAT ATCACACCCC 10320  
AAACAGAAAT GGCTACTAAA AGAGCTCCCT CATAAGG 10357

## (2) INFORMATION FOR SEQ ID NO: 192:

- (1) SEQUENCE CHARACTERISTICS:  
(A) LENGTH: 6867 base pairs  
(B) TYPE: nucleic acid  
(C) STRANDEDNESS: double  
(D) TOPOLOGY: linear

## (xi) SEQUENCE DESCRIPTION: SEQ ID NO: 192:

CGGGACATTC TCAATCTTCT CTCTTTTGT TTTCTCTTCT TTCTATGATA CAATGGAAAA 60  
AATAAATTCA AAAGGAGTTT TTTTATGACT TATCCAAATC TCTTGGACCG CTCTTAACC 120  
TATGTTAAGG TCAACACGCG CTCTGATGAA CACTCTACTA CTACTCCAAG TACACAGAGT 180  
CAGGTTGACT TCGCAACAAA TGCCTAATT CCTGAAATGA AACGTGTTGG ACTGCAAAAT 240  
GTTTACTATC TACCGAATGG TTTTGCTATT GGAACCTTGC CAGCCAACGA TCCGTCTTTA 300  
ACACGTAAGA TTGGTTTAT ATCGCACATG GATACTGCTG ATTTTAATGC TGAAGGAGTC 360  
AATCCACAGG TAATTGAAAA CTACGATGGT GGTGTGATTG AACTAGGGAA TTTGTGTTTC 420  
AAACTCGATC CAGCTGACTT CAAGAGTCTT GAAAAATATC CAGGACAAAC GTCATCACA 480  
ACAGATGGAA CAACCTTGCT AGGTGCTGAT GACAAGTCAG GAATTGCTGA AATTATGACA 540  
GCCATTGAAT ATCTAACTGC TCATCCTGAA ATTAAGCACT GTGAGATTCTG TGTGCTTTT 600  
GGTCCAGATG AAGAAATCGG TGTGCTGCC AATAAATTTG ATGCAGAAGA TTTGATGTG 660  
GATTTTGCTT AACTGTGTA TGGTGGTCCA CTAGGTGAAC TTCAGTACGA GACTTTCTCA 720  
GCCGCTGGTG CTGAATTGCA TTTCCAAGGT CGTAATGTCC ACCCTGGTAC TGCCAAAGGG 780  
CAGATGGTCA ATGCCCTTCA GCTAGCAATT GATTTTCATA ATCAACTTCC AGAAAATGAC 840  
CGACCTGAGT TAACTGAAGG TTACCAAGGT TTTTACCATC TAATGGATGT GACAGGTAGT 900  
GTTGAGGAGG CGCGTGCAAG CTACATCATT CGTGATTTTG AAAAAGATGC CTTTGAAGCG 960  
CGTAAAGCAT CCATGCAATC TATCGCTGAT AAGATGAATG AAGAACTGG GAGCGACCGT 1020  
GTCACTCTCA ACTTGACAGA CCAGTACTAC AATATGAAAG AAGTCATTGA AAAAGATATG 1080  
ACTCCAATTA CCATGCTTAA AGCCGTTATG GAAGATCTAG GTATCACGCC TATTATCGAA 1140

1158

CCAATCCGGG GTGGAACAGA CGGCTCTAAG ATTTCCCTTTA TGGGAATCCC AACTCCGAAT	1200
ATCTTTGCAG GTGGCGAAAA TATGCACGGA CGTTTTGAAT ACGTTAGCCT TCAGACTATG	1260
GAACGTGCAG TTGATACCAT CATTGGCATC GTAGCTTATA AAGGCTAAAA AGACGAGGTA	1320
GCTCAGCTAC TTCGCCTTC TTTTATTCT ACTGCTTTT CTTGATTTC AGTAGTTGTA	1380
GAAGATTCTG TTGTTTCATT TTCTGAAGT GATTCAGCAG GPTTAGAATC TCTGTATTG	1440
CTTGGTTTGT TTTCTGCT AGCAGTTCA ATGTTAGATT CTGCAGTTGC GTTTCGTTGG	1500
TTCTCAGCAC TGGTGTATC ACCATTGCT TCAGCATTC TTGCTGGACT TGTTCCTCA	1560
CTTGGCTAG CTTTGTAGTG GATTTGATGA TTCAAACTA GAATAGCTTT TGTGATTTCA	1620
AGTAAAGCTG TTTTGTCTT ACTCTTAGCA GAAAGTTGAT CTAATAATGC ATCCACCTTA	1680
TCAAAGTCCG CATCAGATCC ATTATTACTT TCTAAATAAG AGTGAAGCGA CATGAGAATA	1740
TCGTAGAGT TTTGATAGAG TACAAOTGTC TGAGGATCTT GCTCAGCATT TTCCTTTCT	1800
TGTTGAAGGG CGCTAGCGAT ACGAGTCAAG ACATCTTTTA CCTGACTGTT TACTTCATCC	1860
AAGTCTGCAT CAGCCTTGT TGTGGCAGCT TTTAGATTT CTACTTCTTC TGCCAAGGAT	1920
TGTCTGATTC CTTCTTCATG GATTGTTC CAGAGTTGAT TTGCCTTGCT CAAAAGACTT	1980
TCTACTTCTT CCTTGCTATC TGTGCGAGAT TATTGGTTGC TATCTACCAT GTACTCCTAA	2040
AACAGGAGAG TTATAATCCA AGATTACAAG GCCTTACAGA AATAAGAAAT CCAGATAAGA	2100
CAATGTTCTG CCAAGACGCT ATTGCTTCG CACAGCAGCA CGGATTCAAT ATGCTTTAAT	2160
TTTAAAGTTT AGGTGTCAAG ACCTCTTTT AGTGTGCCCA AAATTTAGAG AAGTAATCAA	2220
TCAACTAACT TTTATTTTT TCAAACCTTC AGTAACTGA CCTAAAGCTA ACTCAATCTC	2280
TCTTTGTAGA TGCTCTGCT ATCAGCTAGA AGTTGATCTA CTTTGGCCAA GACTGCCTTC	2340
TCATCAAAAG TTCCAGGTTG ATAGTTGGAT TGCAGGGATG GAATCTTGT TTTCAAAGCC	2400
GCTTCATATC CCTTAGTTG AACCTTGATG TAGTGATTGT GGTGCGCATG AGGAATCACA	2460
AAACCTTCTG AATCTTCACT TATAATTCGA TTGGCATCAA AACCATGACC ATCTTCTTCC	2520
TCATGATGGA CATGTAGTGA CGGATTACTT AATACAGAAC TAGAAGAACT TCCTACCTCT	2580
TCCGTGTTAG AGTGTGATGG GGGATTGTTA AGAGATGACT TAGGAATATA GTGATAGTGA	2640
TCCCCATGTC TTAATATATA AGCATCACCT GTATCTCTGA CAATATCATT AGGTTAAAG	2700
ACATATGTGG CTGCTAATTC ACCTGCCGAC AAGTCACTCT CAGGAATGAA ATGATAGTGA	2760
CCACCATGTG GTACTATAGT AGATTGAAAT AGAATATGAG CAAATTGATA AGGGGATTTT	2820
AAAGTAATTT CTAACAATGA TTTAGAACT ATGATGTGCT ATTCTAAATT CAACTCACTA	2880
TATATAACCA TCATCGGTAG TATAACGTCC CTGTAATTTT GCTACAGATA CTCTGCACT	2940

1159

AGCTCCTTTA TCGTCTTTAC CATGTTCTTG TTTTGGCGA TTGATTTCAT CTTTGTTCG	3000
TACATTTTCT GCATGAGCTT GATCTTTAAG GTAAACATAA TACTTTCCAT CTACCTTAAT	3060
AATATATCCT CCCTTAACCT AACTGACGAT ATCTTGATCT TCGGCTGAT AGTTGGGGC	3120
TTTCATTAAAT AGCTCTTCAC TAAAGAGCGC ATCAAAAGGA ACTTTACCAT TATAGTAGTG	3180
ATAATGATCG CCATGAGAAG TTACATAACC TTGATCTGTA ATCTTAATAA CAATTTGTTT	3240
TGCTTGAATT CCTTCTTTTT GACTAACCTA GTCTGGAGTC AAATTTTCAG TCTTCTTAGT	3300
GTCTTTATTA CTGTTTACAT ATGAAACACG ATTTTATCT GTATTGGCCT GTAGCTATG	3360
TTGGTTCAGA GCATAACAC ACAGACTTAA GGAAAGGATA ACAACAGATC CAGCTGCTAT	3420
ATATTTCTTT TTAATTTICA TAATTACCTC ATTTCTATAA TTATTTATAT GATGCTTCA	3480
TTATTTAAATG ATTAATAAAA TTAATTAACC AATTAATTAA CTAGTAAATA TTCCACCTCT	3540
TTTTAAGTTG TATGCAAGA AATTTTATAT ATTAATAATA AAATGAAAT CTCCCAAAGT	3600
CAGAGTTTTA TTTCTAAGTT TTGAGAGAAC TTCATTTTGG ATTCAGACTT TTTCTACTGC	3660
TATTCCTTAC GCTATGAGAT CAGATAAATT CTTTTTTATC ACTTCTCCAC TTGGCAATCT	3720
TAATTCAAATC GTTCCATCCA TATTGAATAT AACACTATCT AAGCCTAATC CGTAACTAGC	3780
TGTAAATTTT TCTAATTTT CTGTACAGG ATCTACTGCT GGAGCTTCCT CTAATGCTGG	3840
ATCTAACATA GGGTCACTCC CCACATTCCT TTCTGGATTC AACATTCAT TATCCGTGA	3900
GTTTTCTGGT TTTACAGTT TTTGTTTGG TGCCCTCTGT AAAGAATCTG CTGGTTTATT	3960
TTCTGTGGT TGGTTCTCAA CTGTCCAGT AGTACTTTT CCATTTTCAG ATGGTTTATT	4020
TTCCACATTT CCTTGAGGTG CTTCTCCTGT AAAATCTGCC ATATTCTTTT TAATGACTTC	4080
TCCCGATGGT AAATATAATT CAATGTTCC GTCCATATTA AACAGACAT TTTCTAGCTT	4140
CATCCCATAA CTTTCAGCAA ATTTTGCTAC TTTTCTTGT ACAGGATCCA CTGTAGGAAC	4200
TTCTTCTAAC GTTGAATTAC TAGTACTATT CCCAGTTCA GAAAGTTTTT CTTTTCTAC	4260
CTTCTCACTA GTCTTTGGTT CTTCTACCTT TTCATCAAGT TTAAAGTTTT CTGTGCTTT	4320
ATTCCTTTTA AATGTGGTA GAATACTGG TTTATCAGT TCATTTTCTT TTTCCAAGAT	4380
AGGTACTTCC ACAATATAAG TCGATTGATT GTCCAAATAA GCATTTGCCA TGAAGGTTAC	4440
AGGAATTTTA TTTCCGCCG TTCTGGTGT TCCTTGGTTT AATTTCCGAA TCGGTAATTT	4500
GATTTCACCA ACTTTATAGT TATTTTCTAA ATAAGCATTT CCATGAAAT CATCAAACAC	4560
TCTGACTAAA GCATCAGTTC CTTTAGGCAC TGCAAATTGA GGCTTCACTC TTAATAAGT	4620
ATCCCTGCA TGGAAAGGAT AGAAAATCGT TTGACTGGCC ATTTTGTAAAG CTAAAGAGGT	4680



1160  
TGGAACTGTA AATGTACCAT CATAACTTAC TTCTGGATAA TCTTTTGAAG CGATAGTATA 4740  
CTTAAATGTT TGTCTGGTA AATAAGGTTG ATCTAATICA AAGTTTGCAA TATTCCTTAC 4800  
TCCTTCTCCA AATACTTTAC CAGATACTTT CTCCAATACT TTCCCATCTG GTGTTATTAA 4860  
TTTACTAGC ATATTGATAC CTAATTTTTT CTCCAATTCA GCGGAAAAC TAAAAGAAAC 4920  
GCGTTTTTGA CCATTGGCTA GAGTAAAGTT TTGATTATTA AACGTACTAT TTTTAAACAA 4980  
ATTAACAACA TTCGTAAATT CTCTCCAGT ATAAACTTTA TTCCCTTCTT TTTTAGCAAC 5040  
TCCTTCTTCG GGTTTAAACA GTTCATAGTT ACTGTGAGAA TGACCAATTC CAACCGGTTT 5100  
ATGTCATCA ATCGGATCTG CATGATGGTG ATCTCCATGC GGATAAATAA TCGCATTTTT 5160  
TTCTTTATTC ACGACAATAC TTTCACGTTT GACACCATAT TGTTCATAA TGCCAGCAAT 5220  
TTTTTCTTCG ATTTTTTTAT CTAATCTTTT CATTTCTTTG GCATTACTTG GATAATCCTG 5280  
TTCATGAGAT GACAAAGAAT CTAATCCATT ATGACTAGTT TTAACCTCCT CTAATGTTT 5340  
TTGCGCASC TAAATTGCTC TTCTGTCAAG TCCTTCTTGA AGAAATAATG ATTGTGCTCT 5400  
CCGTGACTCA TGACAAAACC TGATTCTCTT TCAGCGATAA TACGATTAGC ATCAAATCCG 5460  
TATCCATCTT CTTCATGTTT CTCATGTGAA GTTCTTGGAT TGATTGGAAG AGATGGAGAA 5520  
GGTGTGCTA GACTATTGTT TGGAAGAGTC GGTGCCCCAA TTTGATTGA TTTTGAATG 5580  
TAATGGAAAT GATCACCATG TCTTACAATA TAAGCTGTAG CCGTTTCTTC AACGATATCT 5640  
TTTGGATTAA AAATATAACC ATCAGATGCT GAAGAGAGCT CCTTACTTGT CGPTAAAGAA 5700  
GAAGGATTGC TTGAAAGACT GCCTAGACTA GACACTACTT CATTAGGTTT TGCAATTGTA 5760  
GAAACTGTAG AACCACTCC ACTGATAGGC ACCATTCTGG CAATCTTTTC TTCTAAGGCA 5820  
GAAAGCTTGC TGTAAAGAAAT AAAGTGGTAA TGGTCGCCAT GCGGAATCGC AACTCCATTT 5880  
GGTGATCCAC TGATAATCTT AGCAGGGTCA AAGACCAGGC CATCTGATTC ACTGTAACGT 5940  
TGGCCGCTAG GTGAATCATA GAGTTCCCTC AAAAGACTCT GGAGATTTTC AGATTTATTT 6000  
GCTGCGCTGC TAGTTGATCC TTTTGCTACA GATTGCGTGT TATTGTCACT AGCTGTTGAA 6060  
GAATAGCTTA ACTGACTCGG TTGCATATTT TTCCAGCCA GATGTGCTTT AGCTGCTGCT 6120  
AATCACTAG CAGATAAATC GCTTTTGGGA ATGTAGTGAT AGTGACCTCC ATGAGGAACG 6180  
ATATAAGCAT TACCCGTATC TTCGATAATA TCAGCTGGAT TAAAGACATA ACCATCATTT 6240  
GTCGTATATC GTCCCTGAGA CCTTGCTACA GCAACATTAG AGTTAACCTT CTCATTATCT 6300  
TTGACATGTT CTGTTTTTG ACCATTGATT TCATCTTTAG TTCGAACATT ATCAGCATGA 6360  
GCTGCATCTT TCAGCTAGAC ATAATATTTT CCATCGACCT TGATGATATA ACCACCCCTG 6420  
ACTTCATTGA CAATATCAGC GTCTTTAAGT TGATAGTTTG GATCCTTCAT CAAGAGTTCT 6480

1161

TCACTAAAGA GGGCATCATA AGGAACTTTC CCATTATAGT AATGATAGTG GTCACCGTGT 6540  
 GACGTTACAT AGCCCTGATC TGTAAATTTG ATTACAATTT GCTCAGCCTG AATTCCCTCT 6600  
 TTCTGGCTAA CCTGGTCTGG TGTCAAGTTT TCACCTTTCT GACTTGACTG GCTGCCATCC 6660  
 ACATAAGAGA CACGATTATT GTCCTTATTT TCCTGCCAAC GATGCTGCTT TAGTGCCATAG 6720  
 GCACATAGAC TCAAGGATAC GATAACAGCT GATCCAGCTG CTATATATTT TTTACTAAAT 6780  
 TTCATAAATC CCTCATTTCA ATAAATGATG AAGTTTTC TCAACTTCTT TTACTTTATT 6840  
 AAATAGTTTT CTAAACCCGG GGGTACC 6867

(2) INFORMATION FOR SEQ ID NO: 193:

- (i) SEQUENCE CHARACTERISTICS:
- (A) LENGTH: 999 base pairs
  - (B) TYPE: nucleic acid
  - (C) STRANDEDNESS: double
  - (D) TOPOLOGY: linear

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 193:

CGTTCTAAAA ATGCAGTACG TTTGATTGAG AATCAGTTA AAGGTATGCT TCCACACAAT 60  
 ACACTTGAC GCGCTCAAGG TATGAAGTTG AAAGTATTTG TTGGAGCTGA GCACACTCAC 120  
 GCTGCACAAC AACCAGAAAT TCTTGACATT TCAGGACTTA TCTAAGGAAA GGAACAATAA 180  
 AGTATGTCAC AAGCACAATA TGCAGGTACT GGACGTCGTA AAAACGCTGT TGCACGCGTT 240  
 CGCCTTGTTT CAGGAACTCG TAAATCACT GTTAACAAAA AAGATGTTGA AGAGTACATC 300  
 CCACACGCTG ACCTTCGTCT TGTCAACAAC CAACCATTCG CAGTTACTTC AACTGTAGGT 360  
 TCATACGACG TTTCGTTAA CGTTATAGGT GGTGGATACG CTGOTCAATC AGGAGCTATC 420  
 CGTCACGGTA TCGCTCGTGC CCTTCTTCAA GTAGACCCAG ACTTCCGCGA TTCATTGAAA 480  
 CGCGCAGGAC TTCTTACAGG TGACTCACGT AAAGTTGAAC GTAAGAAACC AGGTCTTAAG 540  
 AAAGCTCGTA AAGCATCACA ATTTAGTAAA CGTTAATTCG AAAGAATTAC TATACTTATA 600  
 CAGAGCACCT TTCGGGGTGT TCTTTTCTTA TACTTTCTTA CTAAATTGGT GCAATTGACA 660  
 CAGTTGTTGC GACTTTAGTC GCTTACAAAT GTGGCTGCAA CCTGACATGG TCAGTTGCCT 720  
 CAAAACGTTA ATCAATACGA TTATATCAAC GTTCAAAGC ACTCAAGGGT TTACCCATAG 780  
 GGTGCTTTTT TCTATACTTT CTAAAAAAGT TTACCCTAAA ATTTGCCCTA AAATTACCCT 840  
 ACTTATTTTT AAGATGTTGG TAGGCAACTT GTCCAGCAGA TAATGGAACT ATGTTTGAAG 900  
 TATTAACATA AGTCTTAGTT GTAACGGTAT CGCTATGAGT TAATGCTTCA GAAATGGCTT 960

727

GCTGCTGGAC TAGCTGCTTC ACCATTGTTT TTAGGATAGT CAGAAATATA GCTTAATTTT 9780  
CCAGTCCATT TTTTATCAGG ATACACTTTA GAAGTAAAGC TTACTTCTTG ACCTACAGAA 9840  
AGGTTGGCTA GATTGTACTC AGACAATTCT CCCTTGACTT GTAAATTTTC ATTGCTGACA 9900  
ATATGAACCA TAACTTGACT CGCCCCGTGT GGAGATTAG AACATTGCT ATTGACTTCG 9960  
AACCACAGTC CCTCTAGGGT ACTGAGAACA GTTGTTCAT CCAATTGACT TTGAGCCTTG 10020  
CTTAATTGGG CCGCAGCATC TGCACGCGCA TCACGGGCAT CACCCAATTG AGCGTCAATA 10080  
GAAGCAACAG AATTCCAGC CACTGGAGTT GGGCTTTCCA CCGTTGCATC TTCTCTCTCT 10140  
ACTGGCGCTG GTAACGTGG AGCCGGAGCT GAAGCGGCTT CATTCGTGC TTGATTGACT 10200  
TCATTGATAT GACGATCTGC CCTAGTACT GCTCGACTAG CTGAATCATA GGCCGCTG 10260  
GCTTCTGAAC TACTGTACTT GACTAAAGCC TGCCCTTCGC TGACCTTATC GCCACAGAA 10320  
ACAAGGATTT CATCTAAATC ACCCTTACTA GCATCAAAT AAACATATTG TTCATTTTTT 10380  
GCTGTACTG TCCCTGACAA TAAACAGAG GAGGCCACGC TTCCTTCCTT GGCAACAACA 10440  
AGATGAGTAG GCTCATCTTT TAGAGCAGTC TGAGAAGGTT GTCTAAAGAG TAAATCCCC 10500  
CCAGCACCCA ATACAACCTAC ACTCGCAGCA CCGATTGCTG CATAAGTTG CCACITTTTA 10560  
GCTTTACCAT TCTTTTCTT CATAATGAAA CTCCTTTTCT TTTTACAAT ACTTTGCTAT 10620  
TATACCAAT TTCCTCCAG CAAACAATC AGTTCAGGAT TAAACAATCG TTCGGAATTT 10680  
TGCTTTTCGG 10690

## (2) INFORMATION FOR SEQ ID NO: 94:

## (1) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 8195 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: double
- (D) TOPOLOGY: linear

## (xi) SEQUENCE DESCRIPTION: SEQ ID NO: 94:

GAGAAAGCGC CCACGTTCC CCGAAGGGAG AAAGGCGGAC AGGTATCCCG TAAGCGGCCA 60  
GGGTCCGAAC AGGAGAGCGC AACGAGGGAG CTTCCCAGGG GGAACGCCT GGTATCTTTA 120  
TAGTCCTGTC GGGTTTCGCC ACCTCTGACT TGAGCGTCGA TTTTGTGAT GCTCGTCAGG 180  
GGGGCGGAGC CTATGGAAA ACGCCAGCAA CGCGGCTTT TTACGGTTCC TGGCCTTTTG 240  
CTGGCCTTTT GCTCACATGT TCTTTCCTGC GTTATCCCCT GATTCTGTGG ATAACCGTAT 300  
TACCGCCTTT GAGTGAGCTG ATACCGCTCG CCGCAGCCGA ACGACCGAGC GCAGCGAGTC 360

728

AGTGAGCGAG GAAGCGGAAG AGCGCCCAAT ACGCAAACCG CCTCTCCCCG CGCGTTGGCC	420
GATTCATTAA TGCAGCTGGC ACGACAGGTT TCCCGACTGG AAAGCGGGCA GTGAGCGCAA	480
CGCAATTAAT GTGAGTTAGC TCACTCATTA GGCACCCCAG GCTTTACACT TTATGCTTCC	540
GGCTCGTATG TTGTGTGGAA TTGTGAGCGG ATAACAATTT CACACAGGAA ACAGCTATGA	600
CGTGATTACG AATTCGAGCT CGGTACCCGG AAAATCCAGA AAATGCTTGA AAAAAATCCT	660
AGAAGATGGT ATAATACTAA ATTGTAAGGG TTATCACATA TAACTCAAAA AAAGAAAGAA	720
CAAAAGGAGA GTCAAACTAT GGCTTCTAAA GATTTCACG TAGTGGCAGA AACAGGTATT	780
CACGCACGTC CAGCAACATT GTTGTACAA ACTGCTAGCA AATTTGCTTC AGATATCACT	840
CTTGAGTACA AAGGTAAATC AGTTAACCTT AAATCAATTA TGGGTGTAT GAGTCTGGT	900
GTTGGCCAAG GTGCTGACGT AACTATCTCA GCTGAAGGTG CAGATGCAGA TGACGCTATC	960
GCTGCAATCT CAGAAACAAT GGAAAAAGAA GGATTGGCAT AAGGGAAATG ACAGAAATGC	1020
TTAAAGGAAT CGCAGCATCT GACGGTGTG CAGTTGCAAA AGCATATCTA CTCGTTCAGC	1080
CGGATTGTGCT ATTTGAGACT ATTACAGTCG AAGATACAAA CGCAGAAGAA GCTCGCCTTG	1140
ATGCCGCTCT ACAGGCATCA CAAGACGAGC TTTCTGTAT TCGCGAGAAA GCAGTAGGTA	1200
CGCTCGGTGA AGAAGCAGCT CAAGTTTTG ATGCTCACTT AATGGTTCCT GCTGACCCAG	1260
AAATGATCAG CCAATCAAG GAACTATCC GTGCGAAGAA AGTGAATGCA GAAGCAGGTC	1320
TGAAAGAAGT TACAGTATG TTTATCACTA TCTTTGAAGG CATGGAAGAC AACCCATACA	1380
TGCAAGAACG CGCAGCGGAT WCCGCGACG TGACAAAACG TGTATTGGCA AACCTTCTTG	1440
GTAAAAAATT GCCAAACCA GCTTCTATCA ATGAAGAGT GATTGTGATT GCGCATGACT	1500
TGACTCCTTC AGATACAGCT CAATTGGACA AAACTTTGT AAAAGCTTTT GTAACCAACA	1560
TTGGTGGACG TACAAGCCAC TCAGCTATCA TGGCAGTAC ACTTGAAAT OCTGCTGTAT	1620
TAGGTACAAA TAACATCACT GAAATCGTTA AAGACGGTGA CATCCTTGCT GTTAACGGGA	1680
TCACTGGAGA AGTGATTATC AACCCAACAG ATGAACAAGC GGCAGAATTT AAGCAGCTG	1740
GTGAAGCCTA TGCGAAACAA AAAGCTGAAT GGGCACTTTT GAAAGATGCT CAAACAGTGA	1800
CTGCTGACGG TAAACACTTC GAGTTGGCTG CTAATATCGG TACTCCAAA GACGTTGAAG	1860
GTGTTAACAA CAACGGTGCA GAAGCTGTTG GACTTTACCG TACAGAGTTC TTGTACATGG	1920
ATTCTCAAGA CTTCCCACT GAAGATGAGC AGTATGAAGC ATACAAGGCT GTTCTTGAAG	1980
CAATGAACGG TAAACCTGTT GTCGTTCTA CAATGGATAT CGGTGGAGAT AAGGAACCTC	2040
CTTACTTCGA TATGCCTCAC GAAATGAACC CATTCCCTGG ATTCCGTGCT CTTCGTATCT	2100
CTATCTCTGA GACTGGAGAT GCTATGTTCC GCACACAAAT CCGTGCTCTT CTTCGTGCT	2160

729

CTGPTCAGG TCAATTGCGT ATCATGTTCC CAATGGTTGC GCTCTTGAAA GAATTCGGTG	2220
CAGCGAAAGC AGTCTTTGAT GAAGAAAAAG CAACCTTCT TGCTGAAGGT GTTGCAAGTTG	2280
CGGATAACAT CCAAGTTGGT ATCATGATCG AGATTCCTGC AGCGGCTATG CTTGCAGACC	2340
AATTTGCTAA AGAAGTTGAC TTCTTCTCAA TTGGTACAAA CGACTTGATC CAATATACAA	2400
TGGCAGCAGA CCGTATGAAC GAACAAGTTT CATACCTTTA CCAACCATAC AACCCATCAA	2460
TCCTACGCTT GATTAACAAT GTGATCAAAG CAGCTCACGC TGAAGGTAAA TGGGCTGGTA	2520
TGTGTGGTGA GATGGCTGGT GACCAACAAG CTGTTCCACT TCTTGTCGGA ATGGGCTTGG	2580
ATGAGTTCTC TATGTCAGCA ACATCTGTAC TTCGTACACG CAGCTTGATG AAGAACTCG	2640
ACACAGCTAA GATGGAAGAG TACGCAAAACC GTGCCCTTAC AGAATGCTCA ACAATGGAAG	2700
AAGTTCTTGA ACTTCAAAAA GAATACGTTA ATTTTGATTA ATCGAAAAGT CCCTGCAACT	2760
CAGITACAGG GATTTTTTTG ATATTTTAAA AAGAATTTTC AAGAAAATCT TTCTTATAGA	2820
AAGTCCAACC TTGAAAAAGT AGTGGTCAGA ACAAAAAATA CTTAAATGGT TCATAAAATT	2880
CTTGACAAGT TGGATATTTA GGAGTAACT ATTAACCACT TAAGTAATAG AGAGGAGTTT	2940
CTGCAATTTA GAAATGAATT GCAACTAGAA ATATCAAATA GAAAGAGAGT TTCGATGAAA	3000
ATTAATAAGA AATACCTTGT TGGTTCTGCG GCACTTTGAT TTAAAGTGTT TGTTCCTACC	3060
AGTTGGGACT GTATCAAGCT AGAACGGTTA AGGAAAAATA TCGTGTTCCT TATATAGATG	3120
GAAACAAGC GACGCAAAAA ACGGAGAATT TGACTCCTGA TGAGGTTAGC AAGCGTGAAG	3180
GAATCAATGC TGAGCAAATC GTCATCAAGA TAACAGACCA AGGCTATGTC ACTTCACATG	3240
GCGACCACTA TCATTATTAC AATGGTAAGG TTCCTTATGA CGCTATCATC AGTGAAGAAT	3300
TACTCATGAA AGATCCAAAC TATAAGCTAA AAGATGAGGA TATTGTTAAT GAGGTCAAGG	3360
GTGGATATGT TATCAAGGTA GATGGAAAAT ACTATGTTTA CCTTAAGGAT GCTGCCACG	3420
CGGATAACGT CCGTACAAAA GAGGAAATCA ATCGACAAAA ACAAGAGCAT AGTCAACATC	3480
GTGAAGGTGG AACTCCAAGA AACGATGGTG CTGTTGCCCTT GGCACGTTCC CAAGGACGCT	3540
ATACTACAGA TGATGTTTAT ATCTTTAATG CTTCTGATAT CATAGAGGAT ACTGGTGATG	3600
CTTATATCGT TCCTCATGGA GATCATTACC ATTACATTCC TAAGAATGAG TTATCAGCTA	3660
GCGAGTTGGC TGCTGCAGAA GCCTTCCTAT CTGGTCGAGG AAATCTGTCA AATTCAAGAA	3720
CCTATCGCCG ACAAATAGC GATAACACTT CAAGAACAAA CTGGGTACCT TCTGTAAGCA	3780
ATCCAGGAAC TACAATACT AACACAAGCA ACAACAGCAA CACTAACAGT CAAGCAAGTC	3840
AAAGTAATGA CATTGATAGT CTCTTGAAAC AGCTCTACAA ACTGCCTTTG AGTCAACGAC	3900

729

CTGTTACGG TCAATTGCGT ATCATGTTCC CAATGGTTGC GCTCTTGAAA GAATTCGGTG	2220
CAGCGAAAGC AGTCTTTGAT GAAGAAAAG CAAACCTTCT TGCTGAAGGT GTTGCAGTTG	2280
CGGATAACAT CCAAGTTGCT ATCATGATCG AGATTCCCTGC AGCGGCTATG CTTGCAGACC	2340
AATTGCTAA AGAAGTTGAC TTCTTCTCAA TTGGTACAAA CGACTTGATC CAATATACAA	2400
TGGCAGCAGA CCGTATGAAC GAACAAGTTT CATACCTTTA CCAACCATAC AACCCATCAA	2460
TCCTACGCTT GATTAACAAT GTGATCAAAG CAGCTCACGC TGAAGGTAAA TGGGCTGGTA	2520
TGTGTGGTGA GATGGCTGGT GACCAACAAG CTGTTCCACT TCTTGTCCGA ATGGGCTTGG	2580
ATGAGTCTCT TATGTCAGCA ACATCTGTAC TTCGTACACG CAGCTTGATG AAGAACTCG	2640
ACACAGCTAA GATGGAAGAG TACGCAAAACC GTGCCCTTAC AGAATGCTCA ACAATGGAAG	2700
AAGTCTTGA ACTTCAAAAA GAATACGTTA ATTTTGATTA ATCGAAAAGT CCCTGCAACT	2760
CAGTTACAGG GATTTTMTTG ATATTTTAAA AAGAATTTT AAGAAAATCT TTCTTATAGA	2820
AAGTCCAACC TTGAAAAAGT AGTGGTCAGA ACAAAAAATA CTTAAATGOT TCATAAAATT	2880
CTTGACAAGT TGGATATTTA GGAGTAACT ATTAACCAGT TAAGTAATAG AGCGAGTTT	2940
CTGCAATTTA GAAATGAATT GCAACTAGAA ATATCAATA GAAAGAGAGT TTCGATGAAA	3000
ATTAATAAGA AATACCTTGT TGGTTCTGCG GCACTTTGAT TTTAAGTGTT TGTCTTACG	3060
AGTTGGGACT GTATCAAGCT AGAACGGTTA AGGAAAATAA TCGTGTTTCC TATATAGATG	3120
GAAAACAAGC GACGCAAAAA ACGGAGAATT TGACTCCTGA TGAGGTTAGC AAGCGTGAAG	3180
GAATCAATGC TGAGCAATC GTCATCAAGA TAACAGACCA AGGCTATGTC ACTTCACATG	3240
GCGACCACTA TCATTATTAC AATGGTAAGG TTCCTTATGA CGCTATCATC AGTGAAGAAT	3300
TACTCTGAA AGATCCAAAC TATAAGCTAA AAGATGAGGA TATTGTTAAT GAGTCAAGG	3360
GTGGATATGT TATCAAGGTA GATGGAAAAT ACTATGTTTA CCTTAAGGAT GCTGCCACG	3420
CGGATAACGT CCGTACAAAA GAGGAAATCA ATCGACAAA ACAAGAGCAT AGTCAACATC	3480
GTGAAGGTGG AACTCCAAGA AACGATGGTG CTGTTGCCTT GGCACGTTTG CAAGGACGCT	3540
ATACTACAGA TGATGTTTAT ATCTTTAATG CTTCTGATAT CATAGAGGAT ACTGGTGATG	3600
CTTATATCGT TCCTCATGGA GATCATTACC ATTACATTCC TAAGAATGAG TTATCAGCTA	3660
GCGAGTTGGC TGCTGCAGAA GCCTTCCTAT CTGGTCGAGG AAATCTCTCA AATTCAAGAA	3720
CCTATCGCCG ACAAATAGC GATAACACTT CAAGAACAAA CTGGGTACCT TCTGTAAGCA	3780
ATCCAGGAAC TACAAATACT AACACAAGCA ACAACAGCAA CACTAACAGT CAAGCAAGTC	3840
AAAGTAATGA CATTGATAGT CTCTTGAAAC AGCTCTACAA ACTGCCCTTG AGTCAACGAC	3900

730

ATGTAGAATC TGATGGCCTT GTCTTTGATC CAGCACAAAT CACAAGTCGA ACAGCTAGAG	3960
GTCTTGCACT GCCACACGGA GATCATTACC ACTTCATCCC TTA CTCTCAA ATGCTGTAAT	4020
TGGAAGAACG AATCGCTCGT ATTATTCCCC TTCGTTATCG TTCAAACCAT TGGGTACCAG	4080
ATTCAGAGCC AGAACACCA AGTCCACAAC CGACTCCGGA ACCTAGTCCA GGCCCGCAAC	4140
CTGCACCAAA TCTTAAATA GACTCAAATT CTCTTTGGT TAGTCAGCTG GTACGAAAAG	4200
TTGGGGAAGG ATATGTATTC GAAGAAAAGG GCATCTCTCG TTATGCTCTT GCGAAAGATT	4260
TACCATCTGA AACTGTTAAA AATCTTGAAA GCAAGTTATC AAAACAAGAG AGTGTTCAC	4320
ACACTTTAAC TGCTAAAAA GAAAATGTTG CTCCTCGTGA CCAAGAATTT TATGATAAAG	4380
CATATAATCT GTTAACGTAG GCTCATAAAG CCTTGTTGA AAATAAGGGT CGTAATTCTG	4440
ATTTCCAAGC CTTAGACAAA TTATTAGAAC GCTTGAATGA TGAATCGACT AATAAAGAAA	4500
AATTGGTAGA TGATTTATTC GCATTCTTAG CACCAATTAC CCATCCAGAG CGACTTGGCA	4560
AACCAAATTC TCAAAATGAG TATACTGAAG ACGAAGTTCG TATTGCTCAA TTAGCTGATA	4620
AGTATACAAC GTCAGATGGT TACATTTTTC ATGAACATGA TATAATCAGT GATGAAGGAG	4680
ATGCATATGT AACGCCTCAT ATGGGCCATA GTCAGTGGAT TGGAAAAGAT AGCCTTTCTG	4740
ATAAGGAAAA AGTTCGAGCT CAAGCCTATA CTAAAGAAAA AGGTATCCTA CCTCCATCTC	4800
CAGACGCAGA TGTTAAAGCA AATCCAAGT GAGATAGTGC AGCAGCTATT TACAATCGTG	4860
TGAAAGGGGA AAAACGAATT CCACTCGTTC GACTTCCATA TATGGTTGAG CATACAGTTG	4920
AGGTAAAAA CGGTAAATTC ATTATTCCTC ATAAGGATCA TTACCATAAT ATTAAATTTG	4980
CTTGGTTTGA TGATCACACA TACAAAGCTC CAAATGGCTA TACCTTGGAA GATTTGTTTG	5040
CGACGATTAA GTACTACGTA GAACACCCTG ACGAACGTCC ACATTCTAAT GATGGATGGG	5100
GCAATGCCAG TGAGCATGTG TTAGGCAAGA AAGACCACAG TGAAGATCCA AATAAGAACT	5160
TCAAAGCGGA TGAAGAGCCA GTAGAGGAAA CACCTGCTGA GCCAGAAGTC CCTCAAGTAG	5220
AGACTGAAAA AGTAGAAGCC CAACTCAAAG AAGCAGAAGT TTTGCTTGGC AAAGTAACGG	5280
ATTCTAGTCT GAAAGCCAAT GCAACAGAAA CTCTAGCTGG TTTACGAAAT AATTGACTC	5340
TTCAAATTAT GGATAACAAT AGTATCATGG CAGAAGCAGA AAAATTACTT GCGTTGTTAA	5400
AAGGAAGTAA TCCTTCATCT GTAAGTAAGG AAAAAATAA CTAATGAAA ATGAAAGTCT	5460
CGATAAAGAG GCTTTCATTT TTATTATGTA TATATGTAAT ATTCTTGACA AGCAATATTA	5520
AAAAGAGTAA ACTATTAACT AGTTAATTAA CCGGTTTATT ACTTTATAGT GAATCAAATA	5580
TACTTAAGAA AAGAGGAAAG AATGAAAATT AATAAAAAAT ATCTAGCAGG TTCAGTGGCA	5640
GTCTTGCCC TAAGTGTTCG TTCCTATGAA CTGCTCGTC ACCAAGCTGG TCAGGTTAAG	5700

731

AAAGAGTCTA ATCGAGTTkC TTATATAGAT GGTGATCAGG CTGGTCAAAA GGCAGAAAAC	5760
TTGACACCAAG ATGAAGTCAG TAAGAGGGAG GGGATCAACG CCGAACAAAT CGTCATCAAG	5820
ATTACGGATC AAGGTTATGT GACCTCTCAT GGAGACCATT ATCATTACTA TAATGGCAAG	5880
GTCCCTTATG ATGCCATCAT CAGTGAAGAG CTCTCATGA AAGATCCGAA TTATCAGTTG	5940
AAGGATTGAG ACATTGTCAA TGAAATCAAG GGTGGTTATG TTATCAAGGT AGATGGAAAA	6000
TACTATGTTT ACCTTAAGGA TGCAGCTCAT GCGGATAATA TTCGGACAAA AGAAGAGATT	6060
AAACGTCAGA AGCAGGAACA CAGTCATAAT CACGGGGGTG GTTCTAACGA TCAAGCAGTA	6120
GTTCGAGCCA GAGCCCAAGG ACGCTATACA ACGGATGATG GTTATATCTT CAATGCATCT	6180
GATATCATG AGGACACGGG TGATGCTTAT ATCGTTCCTC ACGGCGACCA TTACCATTAC	6240
ATTCTAAGA ATGAGTTATC AGCTAGCGAG TTAGCTGCTG CAGAAGCCTA TTGGAATGGG	6300
AAGCAGGGAT CTCGTCCTTC TTCAAGTTCT AGTTATAATG CAAATCCAGC TCAACCAAGA	6360
TTGTCAGAGA ACCACAATCT GACTGTCACT CCAACTTATC ATCAAAATCA AGGGGAAAAC	6420
ATTTCAAGCC TTTTACGTGA ATTGTATGCT AAACCTTAT CAGAACGCCA TGTGGAATCT	6480
GATGGCCTTA TTTTCGACCC AGCGCAAAATC ACAAGTCGAA CCGCCAGAGG TGTAGCTGTC	6540
CCTCATGGTA ACCATTACCA CTTTATCCCT TATGAACAAA TGTCTGAATT GGAAAAACGA	6600
ATTGCTCGTA TTATTCCCTT TCGTTATCGT TCAAACCATT GGTACCAGA TTCAAGACCA	6660
GAACAACCA GTCCACAATC GACTCCGAA CCTAGTCCAA GTCGCGCAAC TGCACCAAT	6720
CCTCAACCAG CTCCAAGCAA TCCAATTGAT GAGAAATTGG TCAAAGAAGC TGTTCGAAAA	6780
GTAGGCGATG GTTATGTCTT TGAGGAGAAT GGAGTTTCTC GTTATATCCC AGCCAAGGAT	6840
CTTTCAGCAG AAACAGCAGC AGGCATTGAT AGCAAACTGG CCAAGCAGGA AAGTTTATCT	6900
CATAAGCTAG GAGCTAAGAA AACTGACCTC CCATCTAGTG ATCGAGAATT TTACAATAAG	6960
GCTTATGACT TACTAGCAAG AATTACCAA GATTTACTTG ATAATAAAGG TCGACAAGTT	7020
GATTTTGAGG CTTTGGATAA CCTGTTGGAA CGACTCAAGG ATGTCYCAAG TGATAAAGTC	7080
AAGTTAGTGG ATGATATTCT TGCCCTCTTA GCTCCGATTC GTCATCCAGA ACGTTTAGGA	7140
AAACCAAATG CGCAATTAC CTACACTGAT GATGAGATTC AAGTAGCCAA GTTGGCAGGC	7200
AAGTACACAA CAGAAOACGG TTATATCTTT GATCCTCGTG ATATAACCAG TGATGAGGGG	7260
GATGCCTATG TAACTCCACA TATGACCCAT AGCCACTGGA TTAATAAAGA TAGTTTGTCT	7320
GAAGCTGAGA GAGCGGCAGC CCAGGCTTAT GCTAAAGAGA AAGGTTTGAC CCCTCCTCG	7380
ACAGACCATC AGGATTGAGG AAATACTGAG GCAAAAGGAG CAGAAGCTAT CTACAACCCG	7440



732

GTGAAAGCAG CTAAGAAGGT GCCACTTGAT CGTATGCCCTT ACAATCTTCA ATATACTGTA	7500
CAAGTCAAAA ACGGTAGTTT AATCATACCT CATATGACC ATTACCATAA CATCAAATTT	7560
GAGTGGTTTG ACGAAGGCCT TTATGAGGCA CCTAAGGGGT ATACTCTTGA GGATCTTTTG	7620
GCGACTGTCA AGTACTATGT CGAACATCCA AACGAACGTC CGCATTGAGA TAATGGTTTT	7680
GGTAACGCTA GCGACCATGT TCGTAAAAAT AAGGTAGACC AAGACAGTAA ACCTGATGAA	7740
GATAAGGAAC ATGATGAAGT AAGTGAGCCA ACTCACCTG AATCTGATGA AAAAGAGAAT	7800
CACGCTGGTT TAAATCCTTC AGCAGATAAT CTTTATAAAC CAAGCACTGA TACGGAAGAG	7860
ACAGAGGAAG AAGCTGAAGA TACCACAGAT GAGGCTGAAA TTCCTCAAGT AGAGAAATCT	7920
GTTATTAAAC CTAAGATAGC AGATGCGGAG GCCTTGCTAG AAAAAGTAAC AGATCCTAGT	7980
ATTAGACAAA ATGCTATGGA GACATTGACT GGTCTAAAA GTAGTCTTCT TCTCGGAACG	8040
AAAGATAATA ACACTATTTC AGCAGAAGTA GATAGTCTCT TGGCTTTGTT AAAAGAAAGT	8100
CAACCGGCTC CTATACAGTA GTAAAAAGAA TGGAGCATAT TTTATGGAGA AGTAACCTTT	8160
CGTGTACTT CTCTTTTTA GAAAAACGTA ACAGA	8195

## (2) INFORMATION FOR SEQ ID NO: 95:

- (i) SEQUENCE CHARACTERISTICS:
- (A) LENGTH: 2004 base pairs
  - (B) TYPE: nucleic acid
  - (C) STRANDEDNESS: double
  - (D) TOPOLOGY: linear

## (xi) SEQUENCE DESCRIPTION: SEQ ID NO: 95:

TTTACTAAAA GGAAAAAGA ACTGATTTCT CAGTCCTTCA TTAATCTTAT TCCACACTAA	60
ATAGGTATGG GTAAACAGGT TGTGACCTT GGTGAATCTC GACTTCAACG TCTTCGAATT	120
CTTCTACCAT TTCTTGAGCG ATTTCAATGG CAAGTCTTC GCTCCGTCT TCACCTACAT	180
AGAAGGTTAC GATTTCACTG TCTTCATCCA ACATATGTTT CAAGGTTTCA GTCAATGTTT	240
GGTGCAATC AGGTTTGAC ACAAGAATTT TTCCATCCAC CATACCTAAA TTATCGTTTT	300
CATGGATTTT TAAGCCATCG ATCGTTGTAT CACGCACGGC TGTGTGACG CTCCCGCTAA	360
CGACATCGCT AAGAGCAGCT GTCATACGCT CTTGGTTTTC TTCAATGGAC TTGCTTGAT	420
CAAAGGCAAG AAGACTTGTG ATACCTTGAG GAAGAGTGGC AGCCTCTACC ACTACCGCTG	480
GTTGCTCCAA AACTTCTGCC GCAGATTGAG CTGCCATGAA GATGTTCTTG TTGTTGGCA	540
AGAAGATGAT GTTACGGGCA TTAACCTGTT CAACAGCCTT GATAAAGTCT TCTGTTGAAG	600
GGTTCATGGT TTGACCGCCT TCGATAACAT AATCCACGCC TTGAGAACAG AAGATATCTG	660

# REPORTS

43. Lysates from frozen brain human tissue were prepared as in (24). Radioactive RT-PCR was performed in a total volume of 50  $\mu$ l containing cDNA synthesized from 0.25  $\mu$ g RNA, 20 mM Tris-HCl, pH 8.4, 50 mM KCl, 1.5 mM MgCl<sub>2</sub>, 0.2 mM dNTPs, 1.7  $\mu$ l [ $\alpha$ -<sup>32</sup>P]CTP, and 0.4  $\mu$ M of the primers as follows: hBDNF5', 5'-AGCCA-GAATCGGAACACCA-3'; hBDNF3', 5'-GCACACCT-GGGTAGGCCAAG-3'. PCR amplification was carried out for 30 cycles. Each cycle consisted of the following steps: 94°C for 30 s, 57°C for 30 s and 72°C for 30 s. The same amount of each cDNA was also amplified, independently, with SNAP-25 (synaptosomal associated protein 25, a presynaptic membrane-associated protein localized in grown cones, axons and presynaptic terminals) specific primers, SNAP-25 5', 5'-CAATGATGCCGAGAAAAT-3'; SNAP25 3', 5'-GGAATCAGCCT-TCTCATTA-3'. PCR products were separated by non-denaturing 8% polyacrylamide gel electrophoresis and visualized by autoradiography. BDNF levels were quantified and normalized relative to SNAP-25 levels.
44. V. O. Ona, et al. *Nature* 399, 263 (1999).
45. Total cellular lysates from conditionally immortalized CNS cells (13, 27) were obtained in a buffer containing Tris 50 mM pH 7.4, 5 mM NaCl, Triton X100 1%, 1 mM DTT, 15 mM EGTA supplemented with 1:100 of Protease Inhibitor Cocktail (Sigma). Immunoprecipitates were obtained by incubating the total cellular lysate (from  $4 \times 10^6$  cells) with Mab2166 (1:1000) following conventional immunoprecipitation protocols and loaded. The blotted proteins were exposed to antibody to Htt Mab2166 (dilution 1:5000; Chemicon, Temecula, CA). RNA was reverse-transcribed into single-stranded cDNA using Superscript II RNase H<sup>-</sup> Reverse Transcriptase (Life Technologies) as described by the manufacturer. PCR was performed in a total volume of 50  $\mu$ l containing 1  $\mu$ g cDNA, 20 mM Tris-HCl, pH 8.4, 50 mM KCl, 1.5 mM MgCl<sub>2</sub>, 0.2 mM dNTPs, 5% dimethyl sulfoxide (DMSO), 0.4  $\mu$ M of Htt-specific primers (5'-CGACCTCGGAAAAGCTGATGAA-3' and 5'-CACACG-GTCTTCTCTGCTACCTGA-3'), 2 U Taq polymerase (Life Technologies). Amplification was carried out for 25 cycles. Each cycle consisted of the following steps: 94°C for 30 s, 56°C for 30 s, and 72°C for 60 s. PCR products were separated by electrophoresis on 2% agarose gel and visualized by staining with ethidium bromide.
46. E. Cattaneo et al., *Trends Neurosci.* 24, 182 (2001).
47. A. C. Bachoud-Levi et al., *Lancet* 356, 1975 (2000).
48. The research described in this manuscript was entirely developed at the Department of Pharmacological Sciences, University of Milano. Supported by grants from the Huntington's Disease Society of America (HDSA, New York), Telethon (Italy #E840) and Ministero dell'Università e della Ricerca Scientifica (Italy, Murst#MM06278849-005), and in part by a grant from the Hereditary Disease Foundation (HDF, Santa Monica) (E.C.) and by funds from Associazione Amici Centro "Dino Ferrari," Milano, Italy (V.S.). T.T. was supported by grants from the Swedish Medical Research Council and Life 2000 Program of the Academy of Finland. We thank R. Molteni for help in setting the RNase Protection Assays. E.C., M.E.M., R.M.F., and M.R.H. are members of the "Coalition for the Cure" (HDSA) and of the "Cure HD Initiative" (HDF).

5 February 2001; accepted 1 June 2001  
Published online 14 June 2001;  
10.1126/science.1059581  
Include this information when citing this paper.

## Complete Genome Sequence of a Virulent Isolate of *Streptococcus pneumoniae*

Hervé Tettelin,<sup>1</sup> Karen E. Nelson,<sup>1</sup> Ian T. Paulsen,<sup>1,2</sup> Jonathan A. Eisen,<sup>1,2</sup> Timothy D. Read,<sup>1</sup> Scott Peterson,<sup>1,3</sup> John Heidelberg,<sup>1</sup> Robert T. DeBoy,<sup>1</sup> Daniel H. Haft,<sup>1</sup> Robert J. Dodson,<sup>1</sup> A. Scott Durkin,<sup>1</sup> Michelle Gwinn,<sup>1</sup> James F. Kolonay,<sup>1</sup> William C. Nelson,<sup>1</sup> Jeremy D. Peterson,<sup>1</sup> Lowell A. Umayam,<sup>1</sup> Owen White,<sup>1</sup> Steven L. Salzberg,<sup>1,4</sup> Matthew R. Lewis,<sup>1</sup> Diana Radune,<sup>1</sup> Erik Holtzapple,<sup>1</sup> Hoda Khouri,<sup>1</sup> Alex M. Wolf,<sup>1</sup> Terry R. Utterback,<sup>1</sup> Cheryl L. Hansen,<sup>1</sup> Lisa A. McDonald,<sup>1</sup> Tamara V. Feldblyum,<sup>1</sup> Samuel Angiuoli,<sup>1</sup> Tanja Dickinson,<sup>1</sup> Erin K. Hickey,<sup>1</sup> Ingeborg E. Holt,<sup>1</sup> Brendan J. Loftus,<sup>1</sup> Fan Yang,<sup>1</sup> Hamilton O. Smith,<sup>1\*</sup> J. Craig Venter,<sup>1\*</sup> Brian A. Dougherty,<sup>5</sup> Donald A. Morrison,<sup>6</sup> Susan K. Hollingshead,<sup>7</sup> Claire M. Fraser.<sup>1,3†</sup>

The 2,160,837-base pair genome sequence of an isolate of *Streptococcus pneumoniae*, a Gram-positive pathogen that causes pneumonia, bacteremia, meningitis, and otitis media, contains 2236 predicted coding regions; of these, 1440 (64%) were assigned a biological role. Approximately 5% of the genome is composed of insertion sequences that may contribute to genome rearrangements through uptake of foreign DNA. Extracellular enzyme systems for the metabolism of polysaccharides and hexosamines provide a substantial source of carbon and nitrogen for *S. pneumoniae* and also damage host tissues and facilitate colonization. A motif identified within the signal peptide of proteins is potentially involved in targeting these proteins to the cell surface of low-guanine/cytosine (GC) Gram-positive species. Several surface-exposed proteins that may serve as potential vaccine candidates were identified. Comparative genome hybridization with DNA arrays revealed strain differences in *S. pneumoniae* that could contribute to differences in virulence and antigenicity.

*Streptococcus pneumoniae* (pneumococcus) has played a pivotal role in the fields of genetics and microbiology. The pioneering studies of Avery, MacLeod, and McCarty in 1944 (1) demonstrated that DNA is the true hereditary material by transforming a noncapsulated, avirulent *S. pneu-*

*moniae* strain with DNA from a capsulated virulent strain. This work highlighted the importance of the bacterial polysaccharide capsule as a key pathogenicity factor.

As a human pathogen, *S. pneumoniae* is the most common bacterial cause of acute respira-

tory infection and otitis media and is estimated to result in over 3 million deaths in children every year worldwide from pneumonia, bacteremia, or meningitis (2). Even more deaths occur among elderly people, among whom *S. pneumoniae* is the leading cause of community-acquired pneumonia and meningitis (3). Since 1990, the number of penicillin-resistant strains has increased from 1 to 5% to 25 to 80% of isolates, and many strains are now resistant to commonly prescribed antibiotics such as penicillin, macrolides, and fluoroquinolones (4).

The complete genome sequence of a capsular serotype 4 isolate of *S. pneumoniae* [designated TIGR4 (5); TIGR indicates The Institute for Genomic Research] was determined by the random shotgun sequencing strategy (6) (GenBank accession number AE005672; see www.tigr.org/tigr-scripts/CMR2/CMRHomePage.spl). This clinical isolate was taken from the blood of a 30-year-old male patient in Kongsvinger, Norway, and is highly invasive and virulent in a mouse model of infection (7).

The genome consists of a single circular chromosome of 2,160,837 base pairs (bp) with a G + C content of 39.7%. Base pair 1 of the chromosome was assigned within the putative origin of replication. Of the 2236 genes identified (8), 1155 are located on the right of the

<sup>1</sup>The Institute for Genomic Research (TIGR), 9712 Medical Center Drive, Rockville, MD 20850, USA.

<sup>2</sup>Johns Hopkins University, Charles and 34th Streets, Baltimore, MD 21218, USA. <sup>3</sup>George Washington University Medical Center, 2300 Eye Street, NW, Washington, DC 20037, USA. <sup>4</sup>Johns Hopkins University, 3400 North Charles Street, Baltimore, MD 21218, USA. <sup>5</sup>Bristol-Myers Squibb PRI, 5 Research Parkway, Wallingford, CT 06492, USA. <sup>6</sup>University of Illinois at Chicago, 900 South Ashland Avenue, Chicago, IL 60607, USA. <sup>7</sup>University of Alabama at Birmingham, 845 19th Street South, Birmingham, AL 35294, USA.

\*Present address: Celera Genomics, 45 West Gude Drive, Rockville, MD 20850, USA.

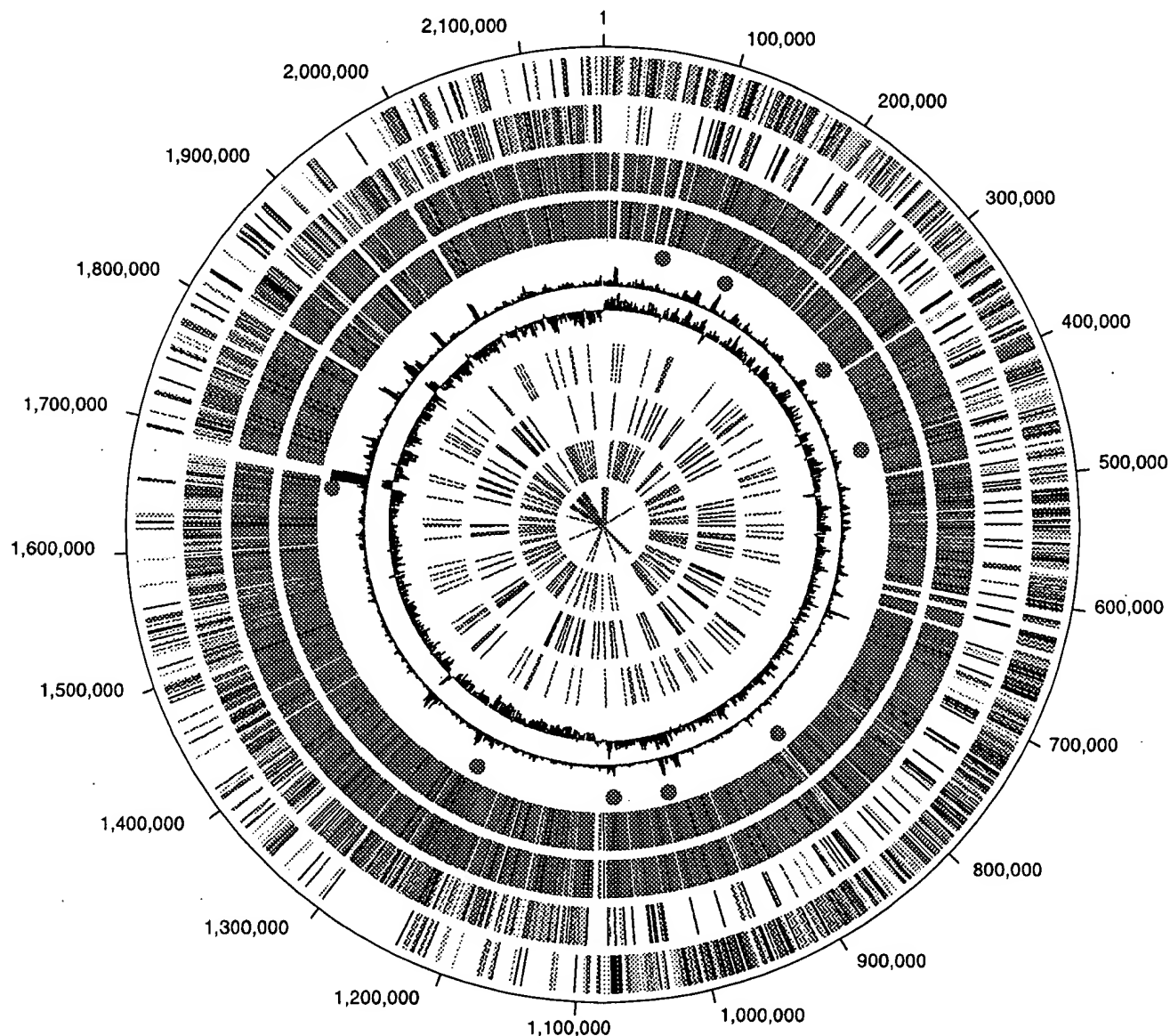
†To whom correspondence should be addressed. E-mail: cmfraser@tigr.org

## REPORTS

origin of replication, and 916 (79%) of these are transcribed in the same direction as DNA replication; similarly, 1081 genes are on the left of the origin of replication, and 857 (79%) of them

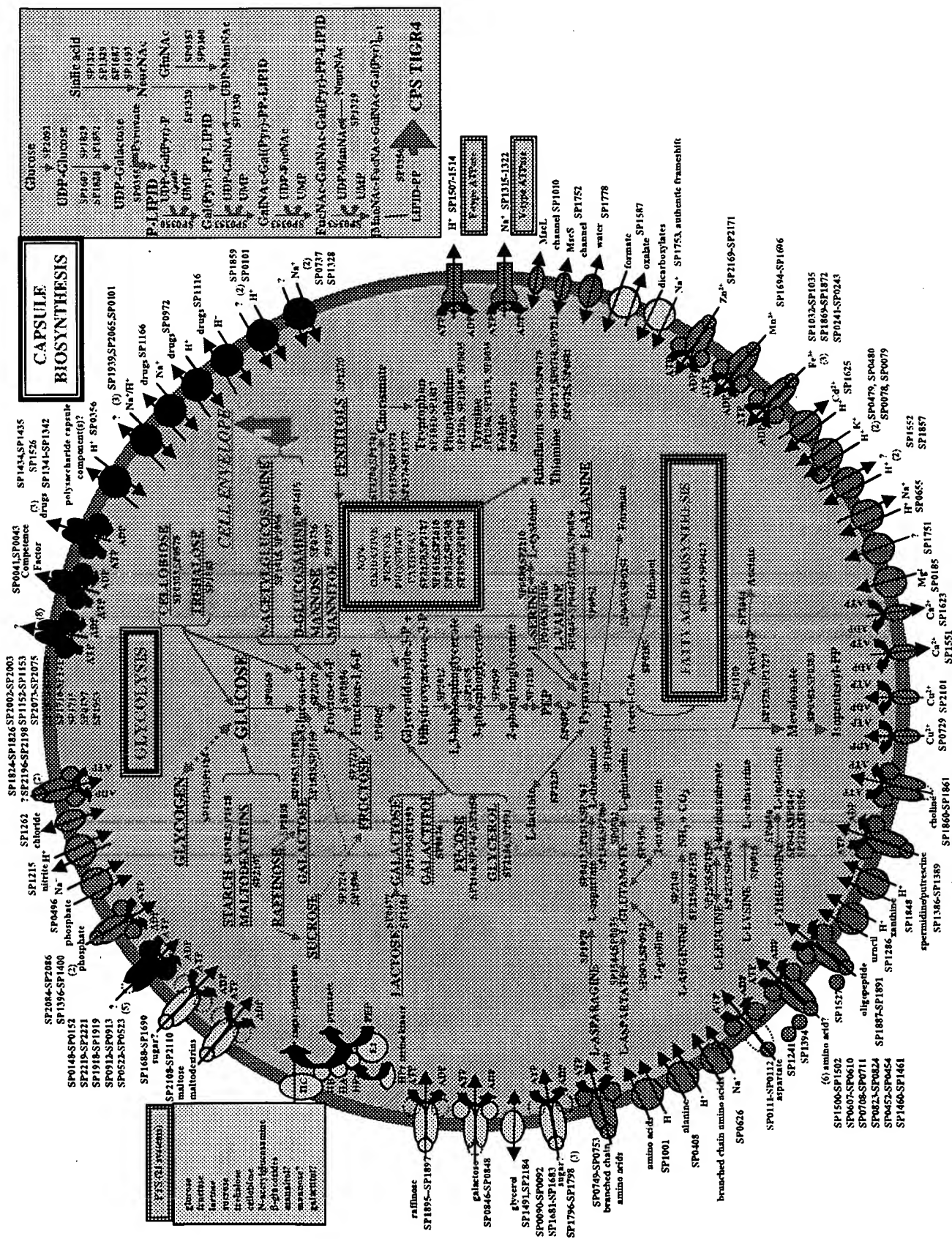
transcribed in the same direction [Fig. 1 and Web fig. 1 (9)]. This type of gene orientation bias appears to be a common feature of low-GC Gram-positive organisms (10).

Although the *S. pneumoniae* genome was reported to contain six ribosomal RNA (rRNA) operons (11), the TIGR4 isolate contains only four rRNA operons. Only 12 of the 58 tRNAs



**Fig. 1.** Circular representation of the *S. pneumoniae* TIGR4 genome and comparative genome hybridizations using microarrays. Comparative genome hybridizations are used to identify genomic differences between the TIGR4 isolate and strains R6 and D39, using a preliminary microarray. Results are displayed on the third and fourth circles. Genes were classified in four groups: (i) gene not present on the array and not analyzed (black) (394 genes, 17% of total); (ii) ortholog present in the test strain (green); (iii) ortholog absent in the test strain (red); and (iv) ambiguous result (blue). The Cy3/Cy5 ratio (TIGR4 signal/test strain) cutoffs for each category were determined subjectively as Cy3/Cy5 = 1.0 to 3.0, green; 3.0 to 10.0, blue; and >10.0, red. There were a number of loci for which hybridization ratios fell between what is expected for gene presence or absence (Cy3/Cy5 ratios between 3.0 to 10.0). Ambiguous results (blue bars) can be explained in at least two ways: (i) The gene may be highly diverged in R6 and/or D39 relative to the TIGR4 isolate. (ii) Alternatively, the gene may be absent in R6 and/or D39 but still be able to produce a hybridization signal, because the TIGR4 isolate gene is a member of a

paralogous gene family or a repetitive element. The outer circle shows predicted coding regions on the plus strand, color-coded by role categories: salmon, amino acid biosynthesis; light blue, biosynthesis of cofactors and prosthetic groups and carriers; light green, cell envelope; red, cellular processes; brown, central intermediary metabolism; yellow, DNA metabolism; green, energy metabolism; purple, fatty acid and phospholipid metabolism; pink, protein fate/synthesis; orange, purines, pyrimidines, nucleosides, and nucleotides; blue, regulatory functions; grey, transcription; teal, transport and binding proteins; black, hypothetical and conserved hypothetical proteins. The second circle shows predicted coding regions on the minus strand, color-coded by role categories. The third circle shows strain R6 genes. The fourth circle shows strain D39 genes. The fifth circle shows an atypical nucleotide composition curve; the nine gene clusters that are absent in strains R6 and D39 are indicated by red bullets. The sixth circle shows the GC-skew curve. The seventh circle shows IS elements. The eighth circle shows RUP elements. The ninth circle shows BOX elements. The tenth circle shows rRNAs in blue, tRNAs in green, and structural RNAs in red.



## REPORTS

are not found adjacent to a rRNA operon [Fig. 1 and Web fig. 1 (9)]. Three structural RNAs were identified: a tRNA-like/mRNA-like (tm) RNA ([www.indiana.edu/~truma/](http://www.indiana.edu/~truma/)), a signal recognition particle RNA (12), and a ribonuclease P RNA (13).

Biological roles were assigned to 1440 (64%) of the predicted proteins according to the classification scheme adapted from Riley (14). Another 359 (16%) predicted proteins matched proteins of unknown function, and the remaining 437 (20%) had no database match. A total of 260 paralogous protein families were identified in the TIGR4 isolate (8), containing 823 predicted proteins (37% of the total).

Comparative genome analysis identified 258 genes in *S. pneumoniae* [Web table 1 (9)] that probably were duplicated after the divergence of this species from other evolutionary lineages for which complete genomes are available (8). Such lineage-specific gene duplications may reveal species-specific adaptations, because gene duplication is frequently accompanied by functional diversification and divergence. These duplications in *S. pneumoniae* include bacteriocin genes, choline-binding proteins, immunoglobulin A (IgA) proteases, immunity proteins, glycosyl transferases, and a large number of hypothetical and conserved hypothetical proteins. Comparison of the complete set of predicted proteins of *S. pneumoniae* with those of other completely sequenced organisms revealed 1219 proteins that are most similar to a protein from another low-GC Gram-positive species (*Lactococcus lactis* has the most with 905) [Web fig. 2 (9)]. Only 105 proteins have no similarity to low-GC Gram-positive proteins [Web table 2 (9)].

Two adjacent genes (SP1467 and SP1468) displayed a high degree of DNA sequence identity (76 and 88%, respectively) between *S. pneumoniae* and *Haemophilus influenzae*. Both pairs of genes, which may be involved in pyridoxine biosynthesis, are more closely related to each other than to orthologs in any other species, which suggests that they were horizontally transferred between these respiratory pathogens.

The *S. pneumoniae* genome is rich in insertion sequences (ISs), which make up ~5% (101,045 bp) of the TIGR4 chromosome [Table 1, Fig. 1, and Web fig. 1 (9)]. IS genes make up

>3.5% (84 out of 2236) of the genes in *S. pneumoniae*, in contrast to other published genomes in which the percentage ranges from 0 to 3% (see [www.tigr.org/tigr-scripts/CMR2/CMRHomePage.spl](http://www.tigr.org/tigr-scripts/CMR2/CMRHomePage.spl)). In addition to IS elements, there are two full-length group II introns and a 1400-bp fragment of the streptococcal conjugative transposon Tn5252. The TIGR4 isolate does not contain any large prophage-like structure or full-length conjugative transposon. The majority of IS elements appear to be non-functional because of insertions, deletions, and/or point mutations (Table 1) that result in frame-shifted or degenerated transposase genes. However, programmed frameshifting may allow the expression of several of the frameshifted genes (15). Intact elements are typically families with 98 to 100% nucleotide sequence identity, probably reflecting "waves" of expansion of IS element isotypes. Despite the large number of IS elements, only two genes (encoding hypothetical proteins SP2178/SP2180 and SP0327/SP0329) are disrupted, and one gene (encoding lacX protein SP1194) is truncated by an IS insertion. This suggests selection against insertions into most of the *S. pneumoniae* genes, or some form of editing to remove these insertions, or both. Regarding the latter, it is possible that the complete DNA transformation system identified in the TIGR4 isolate [Web table 3 (9)] may allow conversion of IS disrupted genes by homologous recombination.

Two types of small, dispersed DNA repeats—the RUP and the BOX elements—were identified previously in *S. pneumoniae*. The 107-bp RUP element is thought to act like a nonautonomous insertion sequence that is mobilized by the transposase of IS630-Spn1 (16). The TIGR4 isolate contains 108 RUP elements, which insert preferentially into IS elements. The BOX element is a modular DNA repeat that is composed of three subunits: *boxA*, *boxB* (which can be present in multiple copies), and *boxC* (17). There are 127 BOX elements in the TIGR4 isolate; of these, 115 are intact ( $A_1B_0C_1$ ) and 12 are incomplete. The BOX elements do not appear to be linked to competence or virulence genes, as was previously suggested (17).

There appears to be a system for generating polymorphic type I restriction enzymes in *S. pneumoniae* similar to that found in *Mycoplas-*

*ma pulmonis* (18). Shotgun sequencing revealed populations of clones from the TIGR4 isolate that were fusions of type I restriction-modification enzyme specificity subunit *hsdS* pseudogenes SP0505 and SP0507 with the nearby intact *hsdS* gene SP0508 [Web fig. 3 (9)]. These rearrangements, which are recombination events between conserved inverted repeats (IRs) within SP0508 and the pseudogenes, might be catalyzed by a nearby integrase (SP0506). Polymerase chain reaction (PCR) on chromosomal DNA using primers inside and outside the *hsdS* genes indicated that the chromosomal region between the IRs was invertible. The specificity subunit may therefore have up to four possible sequences, presumably altering the DNA site recognition of the restriction-modification system and reducing the efficiency of DNA exchange between bacteria in the same clone line.

*Streptococcus pneumoniae* has the widest substrate utilization range for sugars and substituted nitrogen compounds of the three completed genomes of near-commensal residents of the human upper respiratory tract (*H. influenzae*, *Neisseria meningitidis*, and *S. pneumoniae*). Genome analysis suggests that *S. pneumoniae* possesses pathways for catabolism of pentitols via the pentose phosphate pathway, as well as for cellobiose, fructose, fucose, galactose, galactitol, glucose, glycerol, lactose, mannitol, mannose, raffinose, sucrose, trehalose, and maltosaccharides, which can flow directly into the glycolytic pathway (Fig. 2). Ten amino acids and *N*-acetylglucosamine can potentially be used as nitrogen and carbon sources. Genome analysis also revealed a large number of pathways for the complete or partial synthesis of 14 amino acids and chorismate (Fig. 2).

*Streptococcus pneumoniae* contains a high percentage of ATP-dependent transporters, as has been seen in other organisms lacking an electron transfer chain (19). *Streptococcus pneumoniae* possesses both a complete F-type proton adenosine triphosphatase (ATPase) and a V-type ATPase that is probably sodium ion-specific. It also has a sodium ion/proton exchanger and several probable sodium ion-driven transporters (Fig. 2), whose activity would be dependent on the establishment of a sodium motive force. Thus, *S. pneumoniae* can probably interconvert the proton gradient, the sodium

**Fig. 2.** Overview of metabolism and transport in *S. pneumoniae*. Pathways for energy production, metabolism of organic compounds, and capsule biosynthesis are shown. There exist other genes in the capsule biosynthesis locus to which no specific function could be assigned. Transporters are grouped by substrate specificity as follows: inorganic cations (green), inorganic anions (pink), carbohydrates/carboxylates (yellow), amino acids/peptides/amines/purines and pyrimidines (red), and drug efflux and other (black). Question marks indicate uncertainty about the substrate transported. Export or import of solutes is designated by the direction of the arrow through the transporter. The energy-coupling mechanisms of the transporters are also shown: Solutes transported by channel proteins are shown with a double-headed arrow; secondary transporters are shown with two arrowed lines, indicating both the solute and the coupling ion; ATP-driven transporters are

indicated by the ATP hydrolysis reaction; and transporters with an unknown energy coupling mechanism are shown with only a single arrow. Components of transporter systems that function as multisubunit complexes that were not identified are outlined with dotted lines. Where multiple homologous transporters with similar substrate predictions exist, the number of that type of transporter is indicated in parentheses. Systematic gene numbers (SPXXXX) are indicated next to each pathway or transporter; those separated by a dash represent a range of consecutive genes. Details for the PTS transporters are indicated in Web fig. 4 (9). Abbreviations are as follows: ADP, adenosine diphosphate; UMP, uridine monophosphate; UDP, uridine diphosphate; FucNAc, *N*-acetylglucosamine; Gal, galactose; GalNAc, *N*-acetylgalactosamine; GluNAc, *N*-acetylglucosamine; ManNAc, *N*-acetylmannosamine; NeurNAc, *N*-acetylneuraminate; P, phosphate; PP, diphosphate; Pyr, pyruvate.



## REPORTS

ion gradient, and ATP as energy sources, using its F- and V-type ATPases and its sodium ion/proton exchanger. This is somewhat similar to the activity of *Treponema pallidum*, which possesses two V-type ATPases, probably for protons and sodium ions, but no exchanger (20).

Over 30% of the transporters in *S. pneumoniae* were predicted to be sugar transporters (Fig. 2), which is the highest percentage observed to date in any sequenced prokaryote (19). Other completely sequenced respiratory tract organisms, *H. influenzae* and *N. meningitidis*, have a paucity of sugar transporters and are much more reliant on carboxylates and other compounds for their carbon needs. This suggests that *S. pneumoniae* may occupy a distinct microenvironment within the respiratory tract. Host glycoproteins and murein polysaccharides, as well as its own capsular polysaccharides, may be major sources of sugars for *S. pneumoniae*. Reliance on sugar transport and metabolism appears to be a common feature of streptococci, based on their abundance in sugar-rich environments such as the oral cavity (21).

The *S. pneumoniae* sugar transporters primarily consist of phosphoenolpyruvate (PEP)-dependent phosphotransferase system (PTS) transporters and ATP-binding cassette (ABC) transporters. *Streptococcus pneumoniae* has 21 PTS sugar-specific enzyme II complexes with a variety of gene and domain arrangements [Web fig. 4 (9)], more than twice as many as any other sequenced organism relative to genome size, again emphasizing the importance of sugars to the life-style of *S. pneumoniae*. It also possesses single copies of the general PTS enzymes enzyme I and histidine-containing protein (HPr), as well as a HPr serine kinase for regulatory purposes. The *S. pneumoniae* PTS includes systems specific for fructose, glucose, lactose, mannose, mannitol, trehalose, *N*-acetylglucosamine, and sucrose, as well as a variety of PTS systems whose sugar specificities remain to be determined. One PTS system (SP2161 to SP2164) is encoded within a gene cluster including all of the genes necessary for fucose metabolism. *N*-acetylglucosamine is a constituent of the capsule of the TIGR4 isolate, and it is therefore possible that this system may be a PTS for the uptake of *N*-acetylglucosamine or other fucose derivatives. In addition to the PTS, there are seven ABC sugar uptake systems, most of which do not have cytoplasmic ATP-binding components encoded with the other components (Fig. 2).

*Streptococcus pneumoniae* also possesses a variety of ATP- and ion-driven amino acid transporters, as well as transporters for polyamines, uracil, and xanthine. A single ABC transporter lacking a binding protein was found for choline, an important requirement for the streptococcal cell wall. In contrast to the emphasis on sugar transport, only a single transporter was found for monocarboxylates and one for dicarboxylates. *Streptococcus pneumoniae* has a

relatively limited repertoire of transporters for inorganic anion and cations, although this includes a manganese ABC transporter (SP1648

to SP1650) and a zinc transporter (SP2169 to SP2171), which have been associated with virulence (22), as well as three ferric iron and three

Table 1. *S. pneumoniae* IS families.

IS family*	Name (isotype)	IS size (nt)†	Intact transposase	Truncated or frameshifted	Species with homologous elements‡
IS3	IS3-Spn	1359	0	14	Sp Ec My Sg Ne Ha La Ba
IS5	IS1381-Spn	854–860	0	12	La
IS5	IS1515	861	0	1§	Sp Fr Cy La
IS30	IS1239	1046	0	2	Sp So Cl St Ae Le
IS66	IS66	2484–2498	0	7	
IS110	—	?	0	2	
IS605	IS200	747	2	1	Ec Sa Ye En Cl Ha Vi Wo Th De
IS630	IS630-Spn1	896	0	12	Sp Sy Ne
IS1380	IS1380-Spn	1703	11	1	Ab Sp Ba Xa Kl Sm
ISL3	IS1167	1414–1432	8	14	Sp Sh Sd En La St Le MI
Unknown			0	17	
Total			21	84	

\*According to the Mahillon and Chandler classification [J. Mahillon, M. Chandler, *Microbiol. Mol. Biol. Rev.* 62, 725 (1998)]. †Distance between inverted repeats flanking intact or nontruncated IS elements. ‡Species with the most similar elements in GenBank. BlastP hits with an *E* value <10<sup>-20</sup> were included. Key: Ab, *Acetobacter*; Ae, *Aeromonas*; Ba, *Bacillus*; Cl, *Clostridium*; Cy, *Cyanobacterium*; De, *Deinococcus*; Ec, *E. coli*; En, *Enterococcus*; Fr, *Fremyella*; Ha, *Haemophilus*; Kl, *Klebsiella*; La, *Lactobacillus*; Le, *Leuconostoc*; Mi, *Microcystis*; My, *Mycoplasma*; Ne, *Neisseria*; Sa, *Salmonella*; Sg, *S. agalactiae*; Sd, *S. gordonii*; Sh, *S. thermophilus*; Sm, *Sphingomonas*; So, *S. pyogenes*; Sp, *S. pneumoniae*; St, *Staphylococcus*; Sy, *Synechocystis*; Th, *Thermotoga*; Vi, *Vibrio*; Wo, *Wolbachia*; Xa, *Xanthobacter*; Ye, *Yersinia*. §*S. pneumoniae* element demonstrates functional activity [R. Munoz, R. Lopez, E. Garcia, *J. Bacteriol.* 180, 1381 (1998)].

Table 2. Subset of *S. pneumoniae* genes related to virulence containing stretches of iterative DNA that could induce phase-variation. Iterative DNA motifs, including homopolymeric tracts, were searched in the TIGR4 genome [see (29)]. The iterative motifs identified in genes related to virulence are displayed. Abbreviations under "location" are as follows: 5', the motif is in the 5' third of the gene; M, the motif is in the middle third; 3', the motif is in the 3' third; P, the motif is within 50 nt upstream of the translation start site. For SP1772, repeats occur in all three parts of the protein.

ORF	Description	Repeat	Location
SP0071	Immunoglobulin A1 protease	(AT) <sub>4</sub> , (TA) <sub>4</sub>	M, 3'
SP0102	Glycosyl transferase	(G) <sub>6</sub>	M
SP0168	Putative macrolide efflux protein	(TTA) <sub>4</sub>	5'
SP0346	Capsular polysaccharide biosynthesis protein (Cps4A)	(TATT) <sub>3</sub>	5'
SP0349	Capsular polysaccharide biosynthesis protein (Cps4D)	(A) <sub>8</sub>	5'
SP0350	Capsular polysaccharide biosynthesis protein (Cps4E)	(AG) <sub>4</sub>	M
SP0351	Capsular polysaccharide biosynthesis protein (Cps4F)	(A) <sub>8</sub> , (A) <sub>9</sub>	5', 5'
SP0352	Capsular polysaccharide biosynthesis protein (Cps4G)	(AT) <sub>4</sub> , (T) <sub>8</sub>	5', M
SP0353	Capsular polysaccharide biosynthesis protein (Cps4H)	(A) <sub>8</sub>	5'
SP0462	Cell wall surface anchor family protein	(GA) <sub>4</sub>	M
SP0664	Putative zinc metalloprotease (ZmpB)	(CAAAA) <sub>3</sub>	5'
SP0689	UDP- <i>N</i> -acetylglucosamine- <i>N</i> -acetylmuramyl-(pentapeptide) pyrophosphoryl-undecaprenol <i>N</i> -acetylglucosamine transferase	(G) <sub>6</sub> , (G) <sub>6</sub>	5'
SP0907	Putative capsular polysaccharide biosynthesis protein	(G) <sub>6</sub>	5'
SP0966	Adherence and virulence protein A	(A) <sub>8</sub>	5'
SP1267	LicC protein	(ATG) <sub>4</sub> , (AG) <sub>4</sub>	5', M
SP1272	Putative polysaccharide biosynthesis protein	(CT) <sub>4</sub> , (CT) <sub>4</sub>	M, 3'
SP1274	LicD2 protein	(A) <sub>8</sub>	5'
SP1492	Cell wall surface anchor family protein	(CT) <sub>4</sub>	3'
SP1693	Neuraminidase A, authentic frameshift	(T) <sub>8</sub>	5'
SP1769	Glycosyl transferase, authentic frameshift	(C) <sub>9</sub> , (CT) <sub>4</sub>	5', M
SP1772	Cell wall surface anchor family protein	(TCAGCGTCGACAA GTGCGTCGGCC) <sub>540</sub>	
SP1950	Putative bacteriocin formation protein	(T) <sub>9</sub>	P
SP2136	Choline-binding protein (CbpA)	(T) <sub>8</sub> , (T) <sub>8</sub>	5'
SP2145	Antigen, cell wall surface anchor family	(G) <sub>6</sub>	5'
SP2190	Choline-binding protein A (CbpA)	(T) <sub>8</sub> , (T) <sub>8</sub>	5', M

# REPORTS

**Table 3.** *S. pneumoniae* proteins likely to be exposed on the surface, based on computer predictions [see (33)].

ORF	Description	LPxTG*	Choline† binding	Lipoprotein‡	SignalP§	YSIRK	Atypical¶	Repeat#
SP0057	Beta-N-acetylhexosaminidase (StrH)	+			+	+		
SP0069	Choline-binding protein I (Cbpl)		+					
SP0071	Immunoglobulin A1 protease (Iga)	+			+	+		++
SP0082	Cell wall surface anchor family protein	+			+	+		
SP0092	ABC transporter, substrate-binding protein			+	+			
SP0112	Amino acid ABC transporter, periplasmic amino acid-binding protein, putative			+	+			
SP0117	Pneumococcal surface protein A (PspA)		+		+			
SP0148	ABC transporter, substrate-binding protein			+	+			
SP0149	Lipoprotein			+	+			
SP0191	Hypothetical protein			+	+			
SP0198	Hypothetical protein			+	+			
SP0268	Alkaline amylopullulanase, putative	+			+	+		
SP0314	Hyaluronidase	+			+			
SP0368	Cell wall surface anchor family protein, authentic frameshift	+			+	+		
SP0377	Choline-binding protein C (Cbpc)		+		+			
SP0378	Choline-binding protein J (Cbpl)		+		+			
SP0390	Choline-binding protein G (Cbpg)		+					
SP0391	Choline-binding protein F (Cbpf)		+		+			
SP0462	Cell wall surface anchor family protein	+			+			+
SP0463	Cell wall surface anchor family protein	+			+			
SP0464	Cell wall surface anchor family protein	+			+			
SP0468	Sortase, putative			+	+			
SP0498	Endo-beta-N-acetylglucosaminidase, putative	+			+	+		
SP0620	Amino acid ABC transporter, amino acid-binding protein, putative			+	+			
SP0629	Conserved hypothetical protein			+	+			
SP0641	Serine protease, subtilase family	+			+			+++
SP0648	Beta-galactosidase (BgaA)	+			+	+		
SP0659	Thioredoxin family protein			+	+			
SP0664	Zinc metalloprotease ZmpB, putative	+			+		+	+
SP0667	Pneumococcal surface protein, putative		+		+			
SP0771	Peptidyl-prolyl cis-trans isomerase, cyclophilin-type			+	+			
SP0845	Lipoprotein			+	+			
SP0899	Conserved hypothetical protein			+	+			
SP0930	Choline-binding protein E (Cbpe)		+		+			
SP0965	Endo-beta-N-acetylglucosaminidase (LytB)		+		+			
SP0981	Protease maturation protein, putative			+	+			+
SP1000	Thioredoxin family protein			+	+			
SP1002	Adhesion lipoprotein			+	+			
SP1032	Iron-compound ABC transporter, iron compound-binding protein			+	+		+	
SP1154	Immunoglobulin A1 protease (Iga)	+			+	+		
SP1394	Amino acid ABC transporter, amino acid-binding protein			+	+			
SP1400	Phosphate ABC transporter, phosphate-binding protein, putative			+	+			
SP1417	PspC-related protein, degenerate		+					+
SP1492	Cell wall surface anchor family protein	+						+
SP1500	Amino acid ABC transporter, amino acid-binding protein (AatB)			+	+			
SP1527	Oligopeptide ABC transporter, oligopeptide-binding protein (AlbB)			+	+			
SP1573	Lysozyme (LytC)		+		+			+
SP1650	Manganese ABC transporter, manganese-binding adhesion lipoprotein			+	+			
SP1683	Sugar ABC transporter, sugar-binding protein			+	+			
SP1690	ABC transporter, substrate-binding protein			+	+			
SP1772	Cell wall surface anchor family protein	+					+	+(540)
SP1796	ABC transporter, substrate-binding protein			+	+			
SP1826	ABC transporter, substrate-binding protein			+	+			

(Continued on page 504)

# REPORTS

Table 3. (Continued)

ORF	Description	LPxTG*	Choline† binding	Lipoprotein‡	SignalP§	YSIRK	Atypical¶	Repeat#
SP1833	Cell wall surface anchor family protein	+				+	+	
SP1870	Iron-compound ABC transporter, permease protein			+	+			
SP1872	Iron-compound ABC transporter, iron-compound binding protein			+	+			+
SP1891	Oligopeptide ABC transporter, oligopeptide-binding protein (AmiA)			+	+			+
SP1897	Sugar ABC transporter, sugar-binding protein (MsmE)			+				
SP1937	Autolysin (LytA)		+					
SP1975	SpoIIJ family protein			+	+			
SP1992	Cell wall surface anchor family protein	+			+			
SP2041	SpoIIJ family protein			+	+			
SP2084	Phosphate ABC transporter, phosphate-binding protein (PstS)			+	+			
SP2108	Maltose/maltodextrin ABC transporter, maltose/maltodextrin-binding protein (MalX)			+	+			
SP2136	Choline-binding protein (PcpA)		+				+	++
SP2169	Zinc ABC transporter, zinc-binding lipoprotein (AdcA)			+	+			
SP2190	Choline-binding protein A (CbpA)	+	+		+	+		++
SP2197	ABC transporter, substrate-binding protein, putative			+	+			
SP2201	Choline-binding protein D (CbpD)		+		+			

\*Sortase motif. †Choline-binding motif. ‡Lipid attachment motif. §Signal peptide; a Y-score lower limit of 0.3 was used as the cutoff. ||Signal peptide YSIRK for Gram-positive cell wall-attached proteins. ¶ORFs present in regions of atypical nucleotide composition [see (40)]. #ORFs containing iterative DNA motifs that could induce repeat-associated phase variation; one plus sign is shown per motif (exception: SP1772 contains 540 copies of a 24-nt motif).

phosphate ABC transporters. Overcoming iron and phosphate limitation may also be important for virulence. *Streptococcus pneumoniae* possesses an ABC efflux system involved in competence (SP0042 and SP0043). The characterized macrolide efflux proteins MefE and MefA (23) are absent from the TIGR4 isolate.

Analysis of the genome sequence suggests that extracellular enzyme systems for the metabolism of polysaccharides and hexosamines are important for providing carbon and nitrogen for this organism and may be important for the synthesis of the capsule and the virulence of this species. Enzyme systems based on *N*-acetylglucosaminidases,  $\alpha$ - and  $\beta$ -galactosidases, endoglycosidases, hydrolases, hyaluronidases, and neuraminidases are present in *S. pneumoniae*. These enzymes probably enable degradation of host polymers, including mucins, glycolipids, and hyaluronic acid, as well as degradation of the organism's own capsule. These enzymatic activities may serve to increase substrate availability to *S. pneumoniae* by converting larger polymers to products that can be transported into the cell, while at the same time damaging host tissues and facilitating colonization.

Pathogenesis and virulence in *S. pneumoniae* are associated with the inflammation and colonization of host tissues and with bypass of the host immune system [Web table 4 (9)] (24). The polysaccharide capsule is considered to be the primary pneumococcal virulence determinant, allowing for the evasion of the host immune response (25). Although no pathway

has been biochemically characterized for the synthesis of the type 4 capsular polysaccharide, a proposed pathway for capsular biosynthesis derived from the genome analysis is shown in Fig. 2. A 13-gene cluster (SP0346 to SP0360) was identified that is likely to be involved in capsular biosynthesis and secretion. This region of the genome has an atypical nucleotide composition and is flanked by two IS elements on each side. Outside of the IS elements are the *aliA* (also called *plpA*) (SP0366) and *dexB* (SP0342) genes, which also flank the capsule loci in other *S. pneumoniae* strains (26). This gene cluster may not represent the complete pathway for capsular biosynthesis, because several other capsular polysaccharide biosynthesis genes are dispersed elsewhere in the genome. An operon of genes involved in the incorporation of phosphorylcholine into teichoic acid is also present in this genome (SP1267 to SP1274), as are all the genes required for peptidoglycan synthesis.

Phase variation has been described in *S. pneumoniae* and shown to involve variation of multiple cell-surface structures that contribute to the ability of the organism to interact with its host (27). One of the mechanisms involves reversible, high-frequency molecular switching of genes through slippagelike mechanisms at iterative DNA motifs, especially homopolymeric tracts (28). Such motifs were identified in the TIGR4 genome (29), and their location was correlated to predicted genes and their promoters. In total, 397 genes (18%) contain iterative

DNA motifs [Web table 5 (9)] and 25 of these are directly related to virulence (Table 2), including genes from the teichoic acid and capsule pathways that are associated with colony opacity variation (30). In contrast to other pathogenic species, most of the nucleotide repeat-containing genes in *S. pneumoniae* are not frameshifted. This might reflect the presence of general mismatch repair in *S. pneumoniae* (31), a process absent in many pathogens (32).

Sixty-nine proteins that are likely to be exposed on the surface of this organism were identified (Table 3) (33). Genomewide analysis of all predicted signal sequences (34) revealed two discernable clusters. The first cluster contains most of the lipoproteins for which the lipid attachment motif (33) extends beyond the covalently modified cysteine and the membrane-spanning region. This suggests some reuse of lipoprotein signal sequences as evolutionary cassettes. The second cluster, composed of proteins anchored in the cell wall through their sortase motif (33), revealed a previously uncharacterized pentapeptide motif (Y/F)SIRK (35), starting usually at residue 12 (Table 3). A large fraction of the surface proteins of various species of *Streptococcus* and *Staphylococcus* display this motif in their signal peptides. The near-perfect conservation of glycine and serine at the fourth and seventh positions past the pentapeptide, within the predicted transmembrane helix, suggests a specific functional interaction and may reflect a step in cell wall attachment in *S. pneumoniae* and related species.



## REPORTS

Among the newly identified surface-exposed genes are a putative alkaline amylopullulanase (SP0268) and a putative endo- $\beta$ -N-acetylglucosaminidase (SP0498). These two genes could be involved in the degradation of host polysaccharides. Several cell-wall surface anchor family proteins and lipoproteins are also possibly involved in adherence to host cells. An unusual surface-associated component in this genome is a 4776-amino acid protein (SP1772) that contains 540 imperfect repeats of the amino acid motif SASTSASA (35). This protein is similar to the *Lactobacillus brevis* surface layer protein (36) and to proteins from *S. gordonii* and *S. cristatus*. It is adjacent to seven glycosyl transferases (SP1758, SP1764 to SP1767, SP1770, and SP1771) that could make O-linked glycosylations on the serines in SP1772. This would produce a structure similar to mucins that might also coat the surface of the bacterium or interact with host cellular mucins, although some strains of *S. pneumoniae* have been shown not to interact with mucins (37).

Comparative genome hybridizations on DNA microarrays were performed (38) between the TIGR4 isolate and both the R6 noncapsulated laboratory strain and the closely related D39 serotype 2 capsulated strain (39). Nine gene clusters in the TIGR4 isolate did not hybridize with the other two strains [Fig. 1 and Web table 6 (9)], which suggests that they are absent or significantly divergent in strains R6 and D39. Six of these regions display an atypical nucleotide composition [Fig. 1 and Web table 7 (9)] (40), which suggests that they were horizontally acquired by the TIGR4 isolate. These include the capsule biosynthesis locus (SP0347 to SP0353), the V-type ATPase locus (SP1315 to SP1322), a gene cluster encoding a cell wall surface anchor protein (SP1772) and seven glycosyl transferases, and a putative macrolide efflux protein (SP0168). In addition to these regions, strains R6 and D39 also lack three putative sortases and two sortase motif proteins (SP0463 to SP0468), as well as choline-binding protein I (SP0069) and an IgA1 protease paralog (SP0071). Similar differences in the capsule locus, IgA1 protease, and choline-binding protein were identified by Hakenbeck *et al.* (41) by means of an oligonucleotide-based microarray. The majority of the loci that differ between the three strains are surface-exposed and/or related to pathogenesis, and these differences may contribute to differences in virulence and antigenicity between these strains.

The complete genome sequence of *S. pneumoniae* has revealed new insights into the complexity of its biology and metabolism, particularly with regard to the dual role of extracellular enzyme systems to provide essential nutrients while at the same time facilitating the colonization of host tissues. Recent experimental studies based on the preliminary genome sequence of the TIGR4 isolate have revealed new candidate vaccine targets for this species (42). The avail-

ability of the complete genome sequence will provide additional avenues for followup studies on the basic biology and pathogenicity of *S. pneumoniae*.

### References and Notes

- O. T. Avery, C. M. MacLeod, M. McCarty, *J. Exp. Med.* **79**, 137 (1944).
- B. Greenwood, *Philos. Trans. R. Soc. London Ser. B* **354**, 777 (1999).
- D. M. Musher, *Clin. Infect. Dis.* **14**, 801 (1992).
- A. Tomasz, *N. Engl. J. Med.* **333**, 514 (1995); G. V. Doern, A. B. Brueggemann, H. Huynh, E. Wingert, *Emerg. Infect. Dis.* **5**, 757 (1999).
- The TIGR4 isolate was previously referred to as JNR7/87, the label of the clinical isolate [A. L. Bricker, A. Camilli, *FEMS Microbiol. Lett.* **172**, 131 (1999)]; as KNR7/87 [A. de Saizieu *et al.*, *J. Bacteriol.* **182**, 4696 (2000)]; R. Hakenbeck *et al.*, *Infect. Immun.* **69**, 2477 (2001)]; and as N4 [T. M. Witzmann *et al.*, *Infect. Immun.* **69**, 1593 (2001)]. Midway through the sequencing project, it became evident that one particular bacterial stock was contaminated with *S. gordonii*, because reads from libraries made with DNA derived from this stock were composed entirely of non-*S. pneumoniae* sequences (assessed by using all available *S. pneumoniae* and *S. gordonii* sequences in GenBank) and would not assemble with the *S. pneumoniae* DNA. Because all aspects of the sequencing project are tracked through a relational database [R. D. Fleischmann *et al.*, *Science* **269**, 496 (1995)], the problem was addressed by identifying and removing all the reads from the libraries in question from the project [*S. gordonii* sequences are available on TIGR's Web site [www.tigr.org/tdb/s\\_gordonii.shtml](http://www.tigr.org/tdb/s_gordonii.shtml)]. The *S. pneumoniae* single-colony isolate that was grown for use in all subsequent libraries was named TIGR4.
- Cloning, sequencing, and assembly were as described [W. C. Nierman *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **98**, 4136 (2001)]. Four small insert (~1.5 kb) shotgun libraries were constructed in pUC-derived vectors after random mechanical shearing (nebulization) of genomic DNA, and three large insert (~18 kb) shotgun libraries were constructed in  $\lambda$ -DASH II vectors (Stratagene) after partial Sau 3A digestion of genomic DNA. Sequencing of the small insert libraries was achieved at a success rate of 66%, with an average read length of 518 bp. The first library constructed was nonrandom, but improvement of the construction methods provided subsequent random libraries. In contrast, none of the large insert libraries appeared to be completely random. Sequencing of these yielded the following success rates per library: first, 366 nucleotides (nt) average length, with a success rate of 26%; second, 620 nt at 52%; and third, 597 nt at 66%. In the late stages of closure, the newly engineered TIGR vector PHOS2 (a pBR derivative) was used to construct a new large insert (~9 kb) library. Sequencing rates were 508 nt at 48.5% success; these are low values, but the library was substantially more random than the lambda libraries. 40,839 small insert and 3449 large insert end sequences were jointly assembled into 390 contigs larger than 1.5 kb (with 220 sequencing gaps and 170 physical gaps) using TIGR Assembler [G. S. Sutton, O. White, M. D. Adams, A. R. Kerlavage, *Genome Sci. Technol.* **1**, 9 (1995)]. The coverage criteria were that every position required at least double-clone coverage (or sequence from a PCR product amplified from genomic DNA) and either sequence from both strands or with two different sequencing chemistries. The sequence was edited manually with the TIGR Editor, and additional PCR [H. Tettelin, D. Radune, S. Kasif, H. Khouri, S. L. Salzberg, *Genomics* **62**, 500 (1999)] and sequencing reactions were performed to close gaps, improve coverage, and resolve sequence ambiguities. Particularly difficult regions, including SP1772, which contains 540 copies of a 24-bp imperfect repeat, were covered by transposon-assisted sequencing (New England Biolabs pGPS Transposon Kit) and mapping of transposon insertions before assembly.
- I. S. Aaberge, J. Eng, G. Lermark, M. Lovik, *Microb. Pathog.* **18**, 141 (1995).
- Open reading frames (ORFs) likely to encode proteins were predicted by Climmer [S. L. Salzberg, A. L. Delcher, S. Kasif, O. White, *Nucleic Acids Res.* **26**, 544 (1998)]; A. L.
- Delcher, D. Harmon, S. Kasif, O. White, S. L. Salzberg, *Nucleic Acids Res.* **27**, 4636 (1999)]. This program, based on interpolated Markov models, was trained with ORFs larger than 600 bp from the genomic sequence, as well as with the *S. pneumoniae* genes available in GenBank. All predicted proteins larger than 30 amino acids were searched against a nonredundant protein database, as previously described [R. D. Fleischmann *et al.*, *Science* **269**, 496 (1995)]. Frameshifts and point mutations were detected and corrected where appropriate. Remaining frameshifts and point mutations are considered to be authentic and were annotated as "authentic frameshift" or "authentic point mutation." Protein membrane-spanning domains were identified by TopPred [M. G. Claros, G. von Heijne, *Comput. Appl. Biosci.* **10**, 685 (1994)]. The 5' regions of each ORF were inspected to define initiation codons using homologies, position of ribosomal binding sites, and transcriptional terminators. Two sets of hidden Markov models were used to determine ORF membership in families and superfamilies: pfam v5.5 [A. Bateman *et al.*, *Nucleic Acids Res.* **28**, 263 (2000)] and TIGRFAMs 1.0 [D. H. Haft *et al.*, *Nucleic Acids Res.* **29**, 41 (2001)]. Pfam v5.5 hidden Markov models were also used with a constraint of a minimum of two hits to find repeated domains within proteins and mask them. Domain-based paralogous families were then built by performing all-versus-all searches on the remaining protein sequences, using a modified version of a previously described method [W. C. Nierman *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **98**, 4136 (2001)]. The extent of potential lineage-specific gene duplications in this genome was estimated by identification of ORFs that are more similar to other ORFs within the TIGR4 genome than to ORFs from other complete genomes, including those of plasmids, organelles, and phages. All ORFs were searched with FASTA3 against all ORFs from the complete genomes, and matches with a FASTA *p* value of  $10^{-5}$  were considered significant.
- Supplementary Web material is available on Science Online at [www.sciencemag.org/cgi/content/full/293/5529/498/DC1](http://www.sciencemag.org/cgi/content/full/293/5529/498/DC1).
- C. Fraser *et al.*, *Science* **270**, 397 (1995); F. Kunst *et al.*, *Nature* **390**, 249 (1997); L. Banerjee, personal communication; S. Gill, personal communication.
- A. M. Gasc, L. Kauc, P. Barraille, M. Sicard, S. Goodgal, *J. Bacteriol.* **173**, 7361 (1991).
- H. Lutcke, *Eur. J. Biochem.* **228**, 531 (1995).
- N. R. Pace, J. W. Brown, *J. Bacteriol.* **177**, 1919 (1995).
- M. Riley, *Microbiol. Rev.* **57**, 862 (1993).
- M. Chandler, O. Fayet, *Mol. Microbiol.* **7**, 497 (1993).
- M. R. Oggioni, J. P. Claverys, *Microbiology* **145**, 2647 (1999).
- B. Martin *et al.*, *Nucleic Acids Res.* **20**, 3479 (1992).
- K. Dybvig, R. Sitarman, C. T. French, *Proc. Natl. Acad. Sci. U.S.A.* **95**, 13923 (1998).
- L. T. Paulsen, L. Nguyen, M. K. Sliwinski, R. Rabus, M. H. Saier Jr., *J. Mol. Biol.* **301**, 75 (2000).
- C. M. Fraser *et al.*, *Science* **281**, 375 (1998).
- R. G. Quivey, W. L. Kuhnert, K. Hahn, *Adv. Microb. Physiol.* **42**, 239 (2000).
- N. S. Jakubovics, A. W. Smith, H. F. Jenkinson, *Mol. Microbiol.* **38**, 140 (2000); A. Dintilhac, C. Alloing, C. Granadel, J. P. Claverys, *Mol. Microbiol.* **25**, 727 (1997); A. M. Berry, J. C. Paton, *Infect. Immun.* **64**, 5255 (1996).
- M. Santagati, F. Iannelli, M. R. Oggioni, S. Stefani, G. Pozzi, *Antimicrob. Agents Chemother.* **44**, 2585 (2000).
- S. K. Hollingshead, D. E. Briles, *Curr. Opin. Microbiol.* **4**, 71 (2001).
- W. B. Wood, M. R. Smith, *J. Exp. Med.* **90** (1949).
- G. Alloing, P. de Philip, J. P. Claverys, *J. Mol. Biol.* **241**, 44 (1994); J. P. Dillard, M. W. Vandersea, J. Yother, *J. Exp. Med.* **181**, 973 (1995); B. J. Pearce, A. M. Naughton, H. R. Masure, *Mol. Microbiol.* **12**, 881 (1994); E. Garcia, D. Lull, R. Munoz, M. Mollerach, R. Lopez, *Res. Microbiol.* **151**, 429 (2000).
- J. N. Weiser, In *Streptococcus pneumoniae—Molecular Biology and Mechanisms of Disease*, A. Tomasz, Ed. (Mary Ann Liebert, Larchmont, NY, 2000), pp. 245–252.
- N. J. Saunders *et al.*, *Mol. Microbiol.* **37**, 207 (2000).
- Iterative DNA motifs, including homopolymeric tracts, were searched in the TIGR4 genome sequence using the

# REPORTS

REPEATS program [G. Benson, M. S. Waterman, *Nucleic Acids Res.* 22, 4828 (1994)]. The minimum length of homopolymeric tracts was set at eight for A and T and at six for G and C; four tandem copies of di- and trinucleotides; and three copies of tetra-, penta-, and hexanucleotides. Heptanucleotides and above were not found in three or more copies, except for the imperfect repeats in SP1772. The ratio of the observed frequency of homopolymeric tracts to their expected frequency was determined by means of Markov chain analysis, as described [N. J. Saunders et al., *Mol. Microbiol.* 37, 207 (2000)]. It revealed that G or C tracts of 8 bp and A or T tracts of 10 and 11 bp are slightly overrepresented.

30. J. O. Kim et al., *Infect. Immun.* 67, 2327 (1999).
31. O. Humbert, M. Prudhomme, R. Hakenbeck, C. G. Dowson, J. P. Claverys, *Proc. Natl. Acad. Sci. U.S.A.* 92, 9052 (1995).
32. J. A. Eisen, P. C. Hanawalt, *Mutat. Res.* 435, 171 (1999).
33. Putative choline-binding motifs [J. L. Garcia, A. R. Sanchez-Beato, F. J. Medrano, R. Lopez, in *Streptococcus pneumoniae—Molecular Biology and Mechanisms of Disease*, A. Tomasz, Ed. (Mary Ann Liebert, Larchmont, NY, 2000), pp. 231–244] were identified using Pfam hidden Markov model (HMM) PF01473 [A. Bateman et al., *Nucleic Acids Res.* 28, 263 (2000)]. LPXTG-type Gram-positive anchor regions [M. J. Pallen, A. C. Lam, M. Antonio, K. Dunbar, *Trends Microbiol.* 9, 97 (2001)] were detected by Pfam HMM PF00746 and by a new HMM built with HMMER 2.1.1 [S. R. Eddy, *Bioinformatics* 14, 755 (1998)] from a new, curated alignment of the surrounding region in *S. pneumoniae*. Candidate lipoprotein signal peptides [S. Hayashi, H. C. Wu, *J. Bioenerg. Biomembr.* 22, 451 (1990)] were flagged by NH<sub>2</sub>-terminal exact matches to the pattern (DERK)(6)-[LVFMFW-STAG](2)-[LVFMFYTAGCQ]-[AGS]-C (35), culled of hypothetical proteins and cytosolic proteins, aligned manually, and used to generate a new HMM. Proteins matching both the HMM and the regular expression are predicted lipoproteins. Putative signal peptides were identified with SignalP [H. Nielsen, J. Engelbrecht, S. Brunak, G. von Heijne, *Protein Eng.* 10, 1 (1997)].
34. The NH<sub>2</sub>-terminal regions of all proteins predicted to have signal sequences were collected for clustering and alignment with ClustalW and were scrutinized. A HMM based on an edited alignment of 40-residue segments around the (Y/F)SIRK motif found several hundred hits to a nonredundant amino acid database. A more general motif, based on the larger family of YSIRK proteins, is (Y/F)S(A)(V/L)(R/K)(R/K)xxxGxxS (35).
35. Single-letter abbreviations for the amino acid residues are as follows: A, Ala; C, Cys; D, Asp; E, Glu; F, Phe; G, Gly; H, His; I, Ile; K, Lys; L, Leu; M, Met; N, Asn; P, Pro; Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; W, Trp; and Y, Tyr.
36. G. Vidgren, I. Palva, R. Pakkanen, K. Lounatmaa, A. Palva, *J. Bacteriol.* 174, 7419 (1992).
37. J. Davies et al., *Infect. Immun.* 63, 2485 (1995).
38. This method is used to identify genomic differences between the TIGR4 strain and strains R6 and D39. All the predicted genes from the TIGR4 strain were amplified by PCR and arrayed on glass microscope slides as previously described [S. Peterson, R. T. Cline, H. Tettelin, V. Sharov, D. A. Morrison, *J. Bacteriol.* 182, 6192 (2000)]. Genomic DNA for comparative genome hybridization studies was labeled according to protocols provided by J. DeRisi ([www.microarrays.org/pdfs/GenomicDNAlabel\\_B.pdf](http://www.microarrays.org/pdfs/GenomicDNAlabel_B.pdf)), except that genomic DNA was not digested or sheared before labeling. Arrays were scanned with a GenePix 4000B scanner from Axon (Union City, CA), and individual hybridization signals were quantitated with TIGR SPOTFINDER [P. Hegde et al., *Biotechniques* 29, 548 (2000)].
39. M. D. Smith, W. R. Guild, *J. Bacteriol.* 137, 735 (1979).
40. Regions of atypical nucleotide composition were identified by the  $\chi^2$  analysis. The distribution of all 64 trinucleotides (trimers) was computed for the complete genome in all six reading frames, followed by the trimer distribution in 2000-bp windows. Windows overlapped by 1500 bp. For each window, the  $\chi^2$  statistic on the difference between its trimer content and that of the whole genome was computed. The most atypical regions, with a score of 600 and above, were considered in this analysis.

41. R. Hakenbeck et al., *Infect. Immun.* 69, 2477 (2001).
42. T. M. Wlzemann et al., *Infect. Immun.* 69, 1593 (2001).
43. We thank M. Heaney, J. Scott, M. Holmes, V. Sapiro, B. Lee, and B. Vincent for software and database support at TIGR; M. Ermolaeva and M. Perlea for specific computer analyses; the TIGR faculty and sequencing core for expert advice and assistance; I. Aaberge (National Institute of Public Health, Oslo, Norway) for providing the initial

clinical isolate labeled JNR.7/87; and G. Zysk and A. Polissi for sharing specific sequence data not deposited in GenBank. Supported in part by the National Institutes of Allergy and Infectious Diseases (grant R01 AI40645-01A1) and the Merck Genome Research Institute (grant MGRI72).

2 April 2001; accepted 4 June 2001

## NPAS2: An Analog of Clock Operative in the Mammalian Forebrain

Martin Reick,<sup>1</sup> Joseph A. Garcia,<sup>2</sup> Carol Dudley,<sup>1</sup> Steven L. McKnight<sup>1\*</sup>

Neuronal PAS domain protein 2 (NPAS2) is a transcription factor expressed primarily in the mammalian forebrain. NPAS2 is highly related in primary amino acid sequence to Clock, a transcription factor expressed in the suprachiasmatic nucleus that heterodimerizes with BMAL1 and regulates circadian rhythm. To investigate the biological role of NPAS2, we prepared a neuroblastoma cell line capable of conditional induction of the NPAS2:BMAL1 heterodimer and identified putative target genes by representational difference analysis, DNA microarrays, and Northern blotting. Coincidence of NPAS2 and BMAL1 activated transcription of the endogenous *Per1*, *Per2*, and *Cry1* genes, which encode negatively activating components of the circadian regulatory apparatus, and repressed transcription of the endogenous *BMAL1* gene. Analysis of the frontal cortex of wild-type mice kept in a 24-hour light-dark cycle revealed that *Per1*, *Per2*, and *Cry1* mRNA levels were elevated during darkness and reduced during light, whereas *BMAL1* mRNA displayed the opposite pattern. In situ hybridization assays of mice kept in constant darkness revealed that *Per2* mRNA abundance did not oscillate as a function of the circadian cycle in NPAS2-deficient mice. Thus, NPAS2 likely functions as part of a molecular clock operative in the mammalian forebrain.

Locomotor activity, body temperature, endocrine hormones, and metabolic rate fluctuate cyclically with a period of 24 hours. The regulatory apparatus that controls circadian rhythm consists of a transcriptional feedback cycle that is evolutionarily conserved in a wide variety of metazoans (1). In mammals, the activating arm of this cycle is executed by a heterodimeric transcription factor composed of the *Clock* and *BMAL1* gene products (2). The *Clock*:*BMAL1* heterodimer binds directly to regulatory sequences of the genes comprising the negative arm of the transcriptional feedback cycle. The negative components of the regulatory apparatus include three period (*Per*) genes and two cryptochrome (*Cry*) genes (3–11), whose products function in a poorly understood manner to inactivate the *Clock*:*BMAL1* heterodimer. The duration of *Per* and *Cry* activity may be modified by a serine-threonine kinase variously termed casein kinase 1 $\epsilon$  or Tau in mam-

mals and Doubletime in flies (12–14). In the absence of entraining influences, this regulatory apparatus oscillates rhythmically at or near the 24-hour light-dark cycle (i.e., 12 hours light, 12 hours dark). Entrainment derived from light, food, temperature, and metabolic activity can advance or delay the central regulatory apparatus such that it is properly adapted to the summation of these external zeitgebers.

The master pacemaker of circadian rhythm resides in the suprachiasmatic nucleus (SCN), a small group of neurons located at the base of the optic chiasma within the central nervous system (15). Classical transplantation experiments have demonstrated that the SCN is necessary and sufficient to specify circadian rhythm (16, 17). Surprisingly, the same molecular clock is operative in sites peripheral to the SCN (11, 18), including cultured mammalian cells of non-neural origin (19).

Neuronal PAS domain protein 2 (NPAS2, also termed MOP4) is a member of the basic helix-loop-helix (bHLH)-PAS domain family of transcription factors. The gene encoding NPAS2 is expressed in a stereotypic pattern of brain nuclei located within the mammalian forebrain (20, 21). Upon positional cloning of

<sup>1</sup>Department of Biochemistry, <sup>2</sup>Department of Internal Medicine, University of Texas Southwestern Medical Center, 5323 Harry Hines Boulevard, Dallas, TX 75390, USA.

\*To whom correspondence should be addressed. E-mail: smckni@biochem.swmed.edu

## CLUSTAL W (1.74) multiple sequence alignment

```

tr|Q6WNQ5|Q6WNQ5_STRPN      -----CAYALNQHRSQENK-DNNR
tr|Q8CWR4|Q8CWR4_STRR6      -MNQIYLRKEERMKINKKYLKAGSVATLVLSVCAYELGLHQQTVK-ENNR
tr|Q8DPQ2|Q8DPQ2_STRR6      MQLAISNRKRVRSMKINKKYLKAGSVATLVLSVCAYELGLHQQTVK-ENNR
tr|Q9AG74|Q9AG74_STRPN      -----MKINKKYLKAGSVATLVLSVCAYELGLHQQTVK-ENNR
tr|Q9AHT9|Q9AHT9_STRPN      -----MKINKKYLKAGSVATLVLSVCAYELGLHQQTVK-ENNR
tr|Q8DQ08|Q8DQ08_STRR6      -----MKINKKYLKAGSVAVLALSVCSYELGRHQAGQVKKESNR
                                *: * . : : * : . *

tr|Q6WNQ5|Q6WNQ5_STRPN      VSYVDGSQSSQKSENLTDPQVSQKEGIQAEQIVIKITDQGYVTSHGDHYH
tr|Q8CWR4|Q8CWR4_STRR6      VSYIDGKQATQKTENLTPDEVSKREGINAEQIVIKITDQGYVTSHGDHYH
tr|Q8DPQ2|Q8DPQ2_STRR6      VSYIDGKQATQKTENLTPDEVSKREGINAEQIVIKITDQGYVTSHGDHYH
tr|Q9AG74|Q9AG74_STRPN      VSYIDGKQATQKTENLTPDEVSKREGINAEQIVIKITDQGYVTSHGDHYH
tr|Q9AHT9|Q9AHT9_STRPN      VSYIDGKQATQKTENLTPDEVSKREGINAEQIVIKITDQGYVTSHGDHYH
tr|Q8DQ08|Q8DQ08_STRR6      VSYIDGDQAGQKAENLTPDEVSKREGINAEQIVIKITDQGYVTSHGDHYH
                                ***:*. *: *:*****:***:***:*****:*****

tr|Q6WNQ5|Q6WNQ5_STRPN      YYNGKVPYDALFSEELLMKDPNYQLKDADIVNEVKGGYIIKVDGKYYVYL
tr|Q8CWR4|Q8CWR4_STRR6      YYNGKVPYDAIISEELLMKDPNYQLKDEDIISEIKGGYVIKVDGKYYVYL
tr|Q8DPQ2|Q8DPQ2_STRR6      YYNGKVPYDAIIFSEELLMKDPNYKLKDEDIVNEVKGGYVIKVDGKYYVYL
tr|Q9AG74|Q9AG74_STRPN      YYNGKVPYDAIISEELLMKDPNYQLKDEDIISEIKGGYVIKVDGKYYVYL
tr|Q9AHT9|Q9AHT9_STRPN      YYNGKVPYDAIISEELLMKDPNYKLKDEDIVNEVKGGYVIKVDGKYYVYL
tr|Q8DQ08|Q8DQ08_STRR6      YYNGKVPYDAIISEELLMKDPNYQLKDSDIVNEIKGGYVIKVDGKYYVYL
                                *****:*****:*** *: *:*****:*****

tr|Q6WNQ5|Q6WNQ5_STRPN      KDAAHADNVRTKDEINRQKQEHVKDNE---KVNSNVAVARSQGRYTND
tr|Q8CWR4|Q8CWR4_STRR6      KDAAHADNVRTKEEINRQKQEHVQHREGGTPRNDGAVALARSQGRYTDD
tr|Q8DPQ2|Q8DPQ2_STRR6      KDAAHADNVRTKEEINRQKQEHVQHREGGTPRNDGAVALARSQGRYTDD
tr|Q9AG74|Q9AG74_STRPN      KDAAHADNVRTKEEINRQKQEHVQHREGGTPRNDGAVALARSQGRYTDD
tr|Q9AHT9|Q9AHT9_STRPN      KDAAHADNVRTKEEINRQKQEHVQHREGGTPRNDGAVALARSQGRYTDD
tr|Q8DQ08|Q8DQ08_STRR6      KDAAHADNIRTKEEIKRQKQERSHNHN---SRADNAVAAARAQGRYTDD
                                *****:***:***:*****: : : : : : * * *:*****:

tr|Q6WNQ5|Q6WNQ5_STRPN      GYVENPADIIEDTGNAYIVPHRGHYHYIPKSDLSASELAAAKAHLAKG--
tr|Q8CWR4|Q8CWR4_STRR6      GYIFNASDIIEDTGDAYIVPHGDHYHYIPKNELSAASELAAAKAFLSGRGN
tr|Q8DPQ2|Q8DPQ2_STRR6      GYIFNASDIIEDTGDAYIVPHGDHYHYIPKNELSAASELAAAEAFLSGRGN
tr|Q9AG74|Q9AG74_STRPN      GYIFNASDIIEDTGDAYIVPHGDHYHYIPKNELSAASELAAAKAFLSGRGN
tr|Q9AHT9|Q9AHT9_STRPN      GYIFNASDIIEDTGDAYIVPHGDHYHYIPKNELSAASELAAAEAFLSGRGN
tr|Q8DQ08|Q8DQ08_STRR6      GYIFNASDIIEDTGDAYIVPHGDHYHYIPKSDLSASELAAQAQYWNKG--
                                **:*. :*****:***** .*****:*****:*. *:

tr|Q6WNQ5|Q6WNQ5_STRPN      -----NMQP-SQLSYSSTASD---NNTQSVAKGSTSKPANKSEN
tr|Q8CWR4|Q8CWR4_STRR6      LSNRSTYRRQNSDNTSRTNWVPSVSNPGTTNTNTSNNSNTNSQASQSN
tr|Q8DPQ2|Q8DPQ2_STRR6      LSNRSTYRRQNSDNTSRTNWVPSVSNPGTTNTNTSNNSNTNSQASQSN
tr|Q9AG74|Q9AG74_STRPN      LSNRSTYRRQNSDNTSRTNWVPSVSNPGTTNTNTSNNSNTNSQASQSN
tr|Q9AHT9|Q9AHT9_STRPN      LSNRSTYRRQNSDNTSRTNWVPSVSNPGTTNTNTSNNSNTNSQASQSN
tr|Q8DQ08|Q8DQ08_STRR6      -----QGSRPSSSSSHNANPAQPRLSNHNLTVTPTYHQN-QGENI
                                . * . . : : . : . * : : :

tr|Q6WNQ5|Q6WNQ5_STRPN      QSLLELYDPSAQRYSSESDGLVFDPAKIIISRTPNGVAIPHGDHYHFIPY
tr|Q8CWR4|Q8CWR4_STRR6      DSKLKQLYKLPLSQRHVESDGLIFDPAQITSRTANGVAVPHGDHYHFIPY
tr|Q8DPQ2|Q8DPQ2_STRR6      DSKLKQLYKLPLSQRHVESDGLVFDPAQITSRTANGVAVPHGDHYHFIPY
tr|Q9AG74|Q9AG74_STRPN      DSKLKQLYKLPLSQRHVESDGLIFDPAQITSRTANGVAVPHGDHYHFIPY
tr|Q9AHT9|Q9AHT9_STRPN      DSKLKQLYKLPLSQRHVESDGLVFDPAQITSRTANGVAVPHGDHYHFIPY
tr|Q8DQ08|Q8DQ08_STRR6      SSKLRELYAKPLSERHVESDGLIFDPAQITSRTANGVAVPHGDHYHFIPY
                                .***:*. * :*: *****:*****: * * * . * . * . * . * . *

tr|Q6WNQ5|Q6WNQ5_STRPN      SKLSALEEKIARMVPISGTGSTVSTNAKPNEVVSSLGSLSSNPSSLTTSK

```

tr Q8CWR4 Q8CWR4_STRR6	SQLSPLEEKLARIIPLYRSNHWPDSRP-EQSPSQSTPEPSPSQPAPN
tr Q8DPQ2 Q8DPQ2_STRR6	SQMSELEERIARIIPLYRSNHWPDSRP-EQSPSQPTPEPSPGPQPAPN
tr Q9AG74 Q9AG74_STRPN	SQLSPLEEKLARIIPLYRSNHWPDSRP-EQSPSQSTPEPSPSQPAPN
tr Q9AHT9 Q9AHT9_STRPN	SQMSELEERIARIIPLYRSNHWPDSRP-EQSPSQPTPEPSPGPQPAPN
tr Q8DQ08 Q8DQ08_STRR6	SQLSPLEEKLARIIPLYRSNHWPDSRP-EQSPSQSTPEPSPSQPAPN
	*::* ***::**::* : .. :::* * * . : ...*.. :::
tr Q6WNQ5 Q6WNQ5_STRPN	ELSSASDGYIFNPKDIVEETATAYIVRHGDHFHYIPKSNQIGQPTLPNNS
tr Q8CWR4 Q8CWR4_STRR6	PQPAPS-----NP--IDEKLVKEAVRKVG DG--YVFEENGVP R-YIPAKD
tr Q8DPQ2 Q8DPQ2_STRR6	-LKIDS-----N-----SSLVSQLVRKVGE G--YVFEEKGISR-YVFAKD
tr Q9AG74 Q9AG74_STRPN	PQPAPS-----NP--IDEKLVKEAVRKVG DG--YVFEENGVP R-YIPAKD
tr Q9AHT9 Q9AHT9_STRPN	-LKIDS-----N-----SSLVSQLVRKVGE G--YVFEEKGISR-YVFAKD
tr Q8DQ08 Q8DQ08_STRR6	PQPAPS-----NP--IDEKLVKEAVRKVG DG--YVFEENGVP R-YIPAKD
	* * .. : : * : * : : : : : :
tr Q6WNQ5 Q6WNQ5_STRPN	LATPSPSLPINPGTSHEKHEEDGYGFDANRIIAEDES GFVMSHGDNHNYF
tr Q8CWR4 Q8CWR4_STRR6	LSAET---AAGIDSKLAKQESLSHKLGA KK---TD-----LPSSDREFYN
tr Q8DPQ2 Q8DPQ2_STRR6	LPSET---VKNLESKLSKQESVSHTLTAK K---EN-----VAPRDQEFYD
tr Q9AG74 Q9AG74_STRPN	LSAET---AAGIDSKLAKQESLSHKLGA KK---TD-----LPSSDREFYN
tr Q9AHT9 Q9AHT9_STRPN	LPSET---VKNLESKLSKQESVSHTLTAK K---EN-----VAPRDQEFYD
tr Q8DQ08 Q8DQ08_STRR6	LSAET---AAGIDSKLAKQESLSHKLGA KK---TD-----LPSSDREFYN
	*:: : . :. *:. : : *:: : : :. *::*
tr Q6WNQ5 Q6WNQ5_STRPN	FKKDLTTEEQIKAAQKHLEEVKTS HNGLDLSLSHEQDYPSNAKEMKDLDDKK
tr Q8CWR4 Q8CWR4_STRR6	KAYDLLARIHQDILLDN-KGRQVD FEALDNLLERLKDVS SSKVKLV D---D
tr Q8DPQ2 Q8DPQ2_STRR6	KAYNLLTEAHKALFEN-KGRNSDFQALDKLLERLNDE STNKEKLV D---D
tr Q9AG74 Q9AG74_STRPN	KAYDLLARIHQDILLDN-KGRQVD FEALDNLLERLKDVS SSKVKLV D---D
tr Q9AHT9 Q9AHT9_STRPN	KAYNLLTEAHKALFXN-KGRNSDFQALDKLLERLNDE STNKEKLV D---D
tr Q8DQ08 Q8DQ08_STRR6	KAYDLLARIHQDILLDN-KGRQVD FEALDNLLERLKDVS SSKVKLV D---D
	:* . : : : : :. **.* : : * : : : * .
tr Q6WNQ5 Q6WNQ5_STRPN	IEEKIAGIMKQYGVKRESIVVNKEKNAI IYPHGDHHHADPIDEHKPVGIG
tr Q8CWR4 Q8CWR4_STRR6	ILAF LAPIRHP---ER---LGKPNAQIT YTD-----DEIQVAKLAGKY
tr Q8DPQ2 Q8DPQ2_STRR6	LLAF LAPITHP---ER---LGKPNSQIE YTE-----DEVRIAQLADKY
tr Q9AG74 Q9AG74_STRPN	ILAF LAPIRHP---ER---LGKPNAQIT YTD-----DEIQVAKLAGKY
tr Q9AHT9 Q9AHT9_STRPN	LLAF LAPITHP---ER---LGKPNSQIE YTE-----DEVRIAQLADKY
tr Q8DQ08 Q8DQ08_STRR6	ILAF LAPIRHP---ER---LGKPNAQIT YTD-----DEIQVAKLAGKY
	: : * * : : * :. * : * *.. * : : ..
tr Q6WNQ5 Q6WNQ5_STRPN	HSHSNYELFKPEEGVAKKEGNKVYTGEELTNV VNLKLNSTFNNQNFTLAN
tr Q8CWR4 Q8CWR4_STRR6	TTEDGY-IFDPRD-ITSDEGD-AYVTPH MTHSHWIKKDS-LSEAERAAAQ
tr Q8DPQ2 Q8DPQ2_STRR6	TTSDGY-IFDEHD-IISDEGD-AYVTPH MGHSHWIGKDS-LSDKEKVAAQ
tr Q9AG74 Q9AG74_STRPN	TTEDGY-IFDPRD-ITSDEGD-AYVTPH MTHSHWIKKDS-LSEAERAAAQ
tr Q9AHT9 Q9AHT9_STRPN	TTSDGY-IFDEHD-IISDEGD-AYVTPH MGHSHWIGKDS-LSDKEKVAAQ
tr Q8DQ08 Q8DQ08_STRR6	TTEDGY-IFDPRD-ITSDEGD-AYVTPH MTHSHWIKKDS-LSEAERAAAQ
	: ..* :*. : : ..** : *. : : : : * : * : : : . * :
tr Q6WNQ5 Q6WNQ5_STRPN	GQKRVSFSFPPELEKKLGINMLVKLITPDGKVLEKVS GKVFGEGVGNIAN
tr Q8CWR4 Q8CWR4_STRR6	AYAKEKGLTPPSTDH QDSGN-----TEAKGAEAIYNRVKAA-----KK
tr Q8DPQ2 Q8DPQ2_STRR6	AYTKEKGILPPSPDADVKAN-----PTGDSAAAIYNRVKGE-----KR
tr Q9AG74 Q9AG74_STRPN	AYAKEKGLTPPSTDH QDSGN-----TEAKGAEAIYNRVKAA-----KK
tr Q9AHT9 Q9AHT9_STRPN	AYTKEKGILPPSPDADVKAN-----PTGDSAAAIYNRVKGE-----KR
tr Q8DQ08 Q8DQ08_STRR6	AYAKEKGLTPPSTDH QDSGN-----TEAKGAEAIYNRVKAA-----KK
	. : . ** . : . * . . . : : * . .
tr Q6WNQ5 Q6WNQ5_STRPN	FELDQPYLPQGTFKYTIASKDYPEVSYDGTFTVPTSLAYKMASQTIFYPF
tr Q8CWR4 Q8CWR4_STRR6	VPLDR--MP-YNLQYTVEVK-----NGSLIIP---HYDHYHNIKFEWF
tr Q8DPQ2 Q8DPQ2_STRR6	IPLVR--LP-YMVEHTVEVK-----NGNLIIP---HKDHYHNIKFAWF
tr Q9AG74 Q9AG74_STRPN	VPLDR--MP-YNLQYTVEVK-----NGSLIIP---HYDHYHNIKFEWF
tr Q9AHT9 Q9AHT9_STRPN	IPLVR--LP-YMVEHTVEVK-----NGNLIIP---HKDHYHNIKFAWF

```

tr|Q8DQ08|Q8DQ08_STRR6      VPLDR--MP-YNLQYTVEVK-----NGSLIIP---HYDHYHNIKFEWF
. * : : * . : : * * : * : : * : * *

tr|Q6WNQ5|Q6WNQ5_STRPN      HAGDTYLRVNPQFAVPKGTDALVRVFDEFHGNAYLENNYKVGEIKLPIPK
tr|Q8CWR4|Q8CWR4_STRR6      ---DEGLYEAPKGYSLDLLATVKYYVE-HPNERPHSDNGFGNASDHVQR
tr|Q8DPQ2|Q8DPQ2_STRR6      ---DDHTYKAPNGYTLLEDLFATIKYYVE-HPDERPHSDNGWGNASEHVLG
tr|Q9AG74|Q9AG74_STRPN      ---DEGLYEAPKGYSLDLLATVKYYVE-HPNERPHSDNGFGNASDHVQR
tr|Q9AHT9|Q9AHT9_STRPN      ---DDHTYKAPNGYTLLEDLFATIKYYVE-HPDERPHSDNGWGNASEHVLG
tr|Q8DQ08|Q8DQ08_STRR6      ---DEGLYEAPKGYSLDLLATVKYYVE-HPNERPHSDNGFGNASDHVQR
                                *      *:      :. * : : * * :      :. : * : . :

tr|Q6WNQ5|Q6WNQ5_STRPN      LNQGTTTRTAGNKIPVTFMANAYLDNQSTYIVEVPILEKENQTD-----
tr|Q8CWR4|Q8CWR4_STRR6      NKNQGADTNQTEKPNEEKQPTEKPEEETPRECKPQSEKPE-----
tr|Q8DPQ2|Q8DPQ2_STRR6      KKDHSDEPNKNFKADEE-----
tr|Q9AG74|Q9AG74_STRPN      NKNQGADTNQTEKPNEEKQPTEKPEEETPRECKPQSEKPE-----
tr|Q9AHT9|Q9AHT9_STRPN      KKDHSDEPNKNFKADEE-----
tr|Q8DQ08|Q8DQ08_STRR6      NKNQGADTNQTEKPNEEKQPTEKPEEDKEHDEVSEPTHPESDEKENHVGL
                                : :      .      .

tr|Q6WNQ5|Q6WNQ5_STRPN      -----KPSILPQFKRNKAQENSKFDEKVVEPKTSEKVEKEKLSETGN
tr|Q8CWR4|Q8CWR4_STRR6      -P-----KP-----TEEPEESPEES--PEESEEPPQVETEKVKEKLREA--
tr|Q8DPQ2|Q8DPQ2_STRR6      -----P-----VEET--PAEPEVPQVETEKVEAQLKEA--
tr|Q9AG74|Q9AG74_STRPN      -P-----KP-----TEEPEESPEES--PEESEEPPQVETEKVKEKLREA--
tr|Q9AHT9|Q9AHT9_STRPN      -----P-----VEET--PAEPEVPQVETEKVEAQLKEA--
tr|Q8DQ08|Q8DQ08_STRR6      NPSADNLYKPSTDTETEETEEA--EDT--TDEAEIPQVEHSVINAKIAEA--
                                *      *:      : * * : . . : : : * :

tr|Q6WNQ5|Q6WNQ5_STRPN      STSNSTLEEVPVTPVQEKVAKFAESYGMKLENVLFNMDGTIELYLPSGE
tr|Q8CWR4|Q8CWR4_STRR6      ---EDLLGKIQ--NPIIKSNAKETLT-GLK-NNLLFGTQDNNTIMAEA--
tr|Q8DPQ2|Q8DPQ2_STRR6      ---EVLLAKVT--DSSLKANATETLA-GLR-NNLTLQIMDNNSIMAEA--
tr|Q9AG74|Q9AG74_STRPN      ---EDLLGKIQ--NPIIKSNAKETLT-GLK-NNLLFGTQDNNTIMAEA--
tr|Q9AHT9|Q9AHT9_STRPN      ---EVLLAKVT--DSSLKANATETLA-GLR-NNLTLQIMDNNSIMAEA--
tr|Q8DQ08|Q8DQ08_STRR6      ---EALLEKVT--DSSIRQNAVETLT-GLK-SSLLLGTQDNNTISAEV--
                                : * : : . . * : : * : : : : . . :

tr|Q6WNQ5|Q6WNQ5_STRPN      VIKKNMADFTGEAPQGNGENKPSSENGKVSTGTVENQPTENKPADSLPEAP
tr|Q8CWR4|Q8CWR4_STRR6      --EKLLALLKESK-----
tr|Q8DPQ2|Q8DPQ2_STRR6      --EKLLALLKGSNPSSVSKEKIN-----
tr|Q9AG74|Q9AG74_STRPN      --EKLLALLKESK-----
tr|Q9AHT9|Q9AHT9_STRPN      --EKLLALLKGSNPSSVSKEKIN-----
tr|Q8DQ08|Q8DQ08_STRR6      --DSLLALLKESQPTPIQ-----
                                . . : * . . .

tr|Q6WNQ5|Q6WNQ5_STRPN      NEKPVKPENSTDNGMLNPEGNVGSDPMLDPALEEAPAVDPVQEKLEKFTA
tr|Q8CWR4|Q8CWR4_STRR6      -----
tr|Q8DPQ2|Q8DPQ2_STRR6      -----
tr|Q9AG74|Q9AG74_STRPN      -----
tr|Q9AHT9|Q9AHT9_STRPN      -----
tr|Q8DQ08|Q8DQ08_STRR6      -----

tr|Q6WNQ5|Q6WNQ5_STRPN      SYGLGLDSVIFNMDGTIELRLPSGEVIKKNLSDLIA
tr|Q8CWR4|Q8CWR4_STRR6      -----
tr|Q8DPQ2|Q8DPQ2_STRR6      -----
tr|Q9AG74|Q9AG74_STRPN      -----
tr|Q9AHT9|Q9AHT9_STRPN      -----
tr|Q8DQ08|Q8DQ08_STRR6      -----

```

FileUp

MSF: 1086 Type: P Check: 1584 ..

```

Name: tr|Q6WNQ5|Q6WNQ5_STRPN oo Len: 1086 Check: 2031 Weight: 0.100
Name: tr|Q8CWR4|Q8CWR4_STRRR6 oo Len: 1086 Check: 2995 Weight: 0.100
Name: tr|Q8DPQ2|Q8DPQ2_STRRR6 oo Len: 1086 Check: 7473 Weight: 0.100
Name: tr|Q9AG74|Q9AG74_STRPN oo Len: 1086 Check: 1008 Weight: 0.100
Name: tr|Q9AHT9|Q9AHT9_STRPN oo Len: 1086 Check: 5019 Weight: 0.100
Name: tr|Q8DQ08|Q8DQ08_STRRR6 oo Len: 1086 Check: 3058 Weight: 0.100

```

//

```

tr|Q6WNQ5|Q6WNQ5_STRPN ..... .CAYALNQHR SQENK.DNNR
tr|Q8CWR4|Q8CWR4_STRRR6 .MNQIYLRKE ERMKINKKYL AGSVATLVLS VCAYELGLHQ AQTVK.ENNR
tr|Q8DPQ2|Q8DPQ2_STRRR6 MQLEISNRKR VSMKINKKYL VGSAAALILS VCSYELGLYQ ARTVK.ENNR
tr|Q9AG74|Q9AG74_STRPN ..... ..MKINKKYL VGSAAALILS VCSYELGLYQ ARTVK.ENNR
tr|Q9AHT9|Q9AHT9_STRPN ..... ..MKINKKYL VGSAAALILS VCSYELGLYQ ARTVK.ENNR
tr|Q8DQ08|Q8DQ08_STRRR6 ..... ..MKINKKYL AGSVAVLALS VCSYELGRHQ AGQVKKESNR

```

```

tr|Q6WNQ5|Q6WNQ5_STRPN VSYVDGSQSS QKSENLTDPQ VSQKEGIQAE QIVIKITDQG YVTSHGDHYH
tr|Q8CWR4|Q8CWR4_STRRR6 VSYIDGKQAT QKTENLTPDE VSKREGINAE QIVIKITDQG YVTSHGDHYH
tr|Q8DPQ2|Q8DPQ2_STRRR6 VSYIDGKQAT QKTENLTPDE VSKREGINAE QIVIKITDQG YVTSHGDHYH
tr|Q9AG74|Q9AG74_STRPN VSYIDGKQAT QKTENLTPDE VSKREGINAE QIVIKITDQG YVTSHGDHYH
tr|Q9AHT9|Q9AHT9_STRPN VSYIDGKQAT QKTENLTPDE VSKREGINAE QIVIKITDQG YVTSHGDHYH
tr|Q8DQ08|Q8DQ08_STRRR6 VSYIDGDQAG QKAENLTPDE VSKREGINAE QIVIKITDQG YVTSHGDHYH

```

```

tr|Q6WNQ5|Q6WNQ5_STRPN YYNGKVPYDA LFSEELLMKD PNYQLKDADI VNEVKGGYII KVDGKYYVYL
tr|Q8CWR4|Q8CWR4_STRRR6 YYNGKVPYDA IISEELLMKD PNYQLKDEDI ISEIKGGYVI KVDGKYYVYL
tr|Q8DPQ2|Q8DPQ2_STRRR6 YYNGKVPYDA IFSEELLMKD PNYKLKDEDI VNEVKGGYVI KVDGKYYVYL
tr|Q9AG74|Q9AG74_STRPN YYNGKVPYDA IISEELLMKD PNYQLKDEDI ISEIKGGYVI KVDGKYYVYL
tr|Q9AHT9|Q9AHT9_STRPN YYNGKVPYDA IISEELLMKD PNYKLKDEDI VNEVKGGYVI KVDGKYYVYL
tr|Q8DQ08|Q8DQ08_STRRR6 YYNGKVPYDA IISEELLMKD PNYQLKDSDI VNEIKGGYVI KVDGKYYVYL

```

```

tr|Q6WNQ5|Q6WNQ5_STRPN KDAAHADNVR TKDEINRQKQ EHVKDNE... .KVNSNVAVA RSQGRYTTND
tr|Q8CWR4|Q8CWR4_STRRR6 KDAAHADNVR TKEEINRQKQ EHSQHREGGT PRNDGAVALA RSQGRYTTDD
tr|Q8DPQ2|Q8DPQ2_STRRR6 KDAAHADNVR TKEEINRQKQ EHSQHREGGT PRNDGAVALA RSQGRYTTDD
tr|Q9AG74|Q9AG74_STRPN KDAAHADNVR TKEEINRQKQ EHSQHREGGT PRNDGAVALA RSQGRYTTDD
tr|Q9AHT9|Q9AHT9_STRPN KDAAHADNVR TKEEINRQKQ EHSQHREGGT PRNDGAVALA RSQGRYTTDD
tr|Q8DQ08|Q8DQ08_STRRR6 KDAAHADNIR TKEEIKRQKQ ERSNHNN... SRADNAVAAA RAQGRYTTDD

```

```

tr|Q6WNQ5|Q6WNQ5_STRPN GYVFNPADII EDTGNAYIVP HRGHYHYIPK SDLSASELAA AKAHLAGK..
tr|Q8CWR4|Q8CWR4_STRRR6 GYIFNASDII EDTGDAYIVP HGDHYHYIPK NELSASELAA AKAFLSGRGN
tr|Q8DPQ2|Q8DPQ2_STRRR6 GYIFNASDII EDTGDAYIVP HGDHYHYIPK NELSASELAA AEAFLSGRGN
tr|Q9AG74|Q9AG74_STRPN GYIFNASDII EDTGDAYIVP HGDHYHYIPK NELSASELAA AKAFLSGRGN
tr|Q9AHT9|Q9AHT9_STRPN GYIFNASDII EDTGDAYIVP HGDHYHYIPK NELSASELAA AEAFLSGRGN
tr|Q8DQ08|Q8DQ08_STRRR6 GYIFNASDII EDTGDAYIVP HGDHYHYIPK SDLSASELAA AQAYWNGK..

```

```

tr|Q6WNQ5|Q6WNQ5_STRPN ..... NMQP.SQLSY SSTASD...N NTQSVAKGST SKPANKSEN
tr|Q8CWR4|Q8CWR4_STRRR6 LNSRSTYRRQ NSDNTSRTNW VPSVSNPGTT NTNTSNNST NSQASQSN
tr|Q8DPQ2|Q8DPQ2_STRRR6 LNSRSTYRRQ NSDNTSRTNW VPSVSNPGTT NTNTSNNST NSQASQSN

```

tr Q9AG74 Q9AG74_STRPN	LSNSRITYRRQ	NSDNTSRTNW	VPSVSNPGTT	NTNTSNNSNT	NSQASQSNDI
tr Q9AHT9 Q9AHT9_STRPN	LSNSRITYRRQ	NSDNTSRTNW	VPSVSNPGTT	NTNTSNNSNT	NSQASQSNDI
tr Q8DQ08 Q8DQ08_STRR6	.....Q	GSRPSSSSSH	NANPAQPRLS	ENHNLTVTPT	YHQN.QGENI
tr Q6WNQ5 Q6WNQ5_STRPN	QSLLELYDS	PSAQRYSERD	GLVFDPAKII	SRTPNGVAIP	HGDHYHFIPY
tr Q8CWR4 Q8CWR4_STRR6	DSLLKQLYKL	PLSQRHVESD	GLIFDPAQIT	SRTANGVAVP	HGDHYHFIPY
tr Q8DPQ2 Q8DPQ2_STRR6	DSLLKQLYKL	PLSQRHVESD	GLVFDPAQIT	SRTANGVAVP	HGDHYHFIPY
tr Q9AG74 Q9AG74_STRPN	DSLLKQLYKL	PLSQRHVESD	GLIFDPAQIT	SRTANGVAVP	HGDHYHFIPY
tr Q9AHT9 Q9AHT9_STRPN	DSLLKQLYKL	PLSQRHVESD	GLVFDPAQIT	SRTANGVAVP	HGDHYHFIPY
tr Q8DQ08 Q8DQ08_STRR6	SSLLRELYAK	PLSERHVESD	GLIFDPAQIT	SRTANGVAVP	HGDHYHFIPY
tr Q6WNQ5 Q6WNQ5_STRPN	SKLSALEEKI	ARMVPISGTG	STVSTNAKPN	EVVSSLGSL	SNPSSLTTSK
tr Q8CWR4 Q8CWR4_STRR6	SQLSPLEEKI	ARIIPLYRS	NHWVPDSRP	EQPSPQSTPE	PSPSPQAPN
tr Q8DPQ2 Q8DPQ2_STRR6	SQMSELEERI	ARIIPLYRS	NHWVPDSRP	EQPSPQSTPE	PSPSPQAPN
tr Q9AG74 Q9AG74_STRPN	SQLSPLEEKI	ARIIPLYRS	NHWVPDSRP	EQPSPQSTPE	PSPSPQAPN
tr Q9AHT9 Q9AHT9_STRPN	SQMSELEERI	ARIIPLYRS	NHWVPDSRP	EQPSPQSTPE	PSPSPQAPN
tr Q8DQ08 Q8DQ08_STRR6	SQLSPLEEKI	ARIIPLYRS	NHWVPDSRP	EQPSPQSTPE	PSPSPQAPN
tr Q6WNQ5 Q6WNQ5_STRPN	ELSSASDGYI	FNPKDIVEET	ATAYIVRHGD	HFHYIPKSNQ	IGQPTLPNNS
tr Q8CWR4 Q8CWR4_STRR6	PQPAPS....	.NP..IDEKL	VKEAVRKVD	G..YVFEENG	VPR.YIPAKD
tr Q8DPQ2 Q8DPQ2_STRR6	.LKIDS....	.N....SSL	VSQVLRKVD	G..YVFEENG	ISR.YVFAKD
tr Q9AG74 Q9AG74_STRPN	PQPAPS....	.NP..IDEKL	VKEAVRKVD	G..YVFEENG	VPR.YIPAKD
tr Q9AHT9 Q9AHT9_STRPN	.LKIDS....	.N....SSL	VSQVLRKVD	G..YVFEENG	ISR.YVFAKD
tr Q8DQ08 Q8DQ08_STRR6	PQPAPS....	.NP..IDEKL	VKEAVRKVD	G..YVFEENG	VPR.YIPAKD
tr Q6WNQ5 Q6WNQ5_STRPN	LATPSPSLPI	NPGTSHEKHE	EDGYGFDANR	IIAEDESGFV	MSHGDHNNHYF
tr Q8CWR4 Q8CWR4_STRR6	LSAET...AA	GIDSKLAKQE	SLSHKLGAKK	...TD.....	LPSSDREFYN
tr Q8DPQ2 Q8DPQ2_STRR6	LPSET...VK	NLESKLSKQE	SVSHTLTAKK	...EN.....	VAPRDQEFYD
tr Q9AG74 Q9AG74_STRPN	LSAET...AA	GIDSKLAKQE	SLSHKLGAKK	...TD.....	LPSSDREFYN
tr Q9AHT9 Q9AHT9_STRPN	LPSET...VK	NLESKLSKQE	SVSHTLTAKK	...EN.....	VAPRDQEFYD
tr Q8DQ08 Q8DQ08_STRR6	LSAET...AA	GIDSKLAKQE	SLSHKLGAKK	...TD.....	LPSSDREFYN
tr Q6WNQ5 Q6WNQ5_STRPN	FKKDLTEEQI	KAAQKHLEEV	KTSHNGLDL	SSHEQDYPSN	AKEMKDLDKK
tr Q8CWR4 Q8CWR4_STRR6	KAYDLLARIH	QDLLDN.KGR	QVDFEALDNL	LERLKDVS	KVKLVD...D
tr Q8DPQ2 Q8DPQ2_STRR6	KAYNLLTEAH	KALFEN.KGR	NSDFQALDKL	LERLNDESTN	KEKLVD...D
tr Q9AG74 Q9AG74_STRPN	KAYDLLARIH	QDLLDN.KGR	QVDFEALDNL	LERLKDVS	KVKLVD...D
tr Q9AHT9 Q9AHT9_STRPN	KAYNLLTEAH	KALFXN.KGR	NSDFQALDKL	LERLNDESTN	KEKLVD...D
tr Q8DQ08 Q8DQ08_STRR6	KAYDLLARIH	QDLLDN.KGR	QVDFEALDNL	LERLKDVS	KVKLVD...D
tr Q6WNQ5 Q6WNQ5_STRPN	IEEKIAGIMK	QYGVKRESIV	VNKEKNAIY	PHGDHHHADP	IDEHKPVGIG
tr Q8CWR4 Q8CWR4_STRR6	ILAFLAPIRH	P...ER....	LGKPNQITY	TD.....DE	IQVAKLAGKY
tr Q8DPQ2 Q8DPQ2_STRR6	LLAFLAPITH	P...ER....	LGKPNQIEY	TE.....DE	VRIAQLADKY
tr Q9AG74 Q9AG74_STRPN	ILAFLAPIRH	P...ER....	LGKPNQITY	TD.....DE	IQVAKLAGKY
tr Q9AHT9 Q9AHT9_STRPN	LLAFLAPITH	P...ER....	LGKPNQIEY	TE.....DE	VRIAQLADKY
tr Q8DQ08 Q8DQ08_STRR6	ILAFLAPIRH	P...ER....	LGKPNQITY	TD.....DE	IQVAKLAGKY
tr Q6WNQ5 Q6WNQ5_STRPN	HSHSNYELFK	PEEGVAKKEG	NKVYTGEELT	NVVNLLKNST	FNNQNFTLAN
tr Q8CWR4 Q8CWR4_STRR6	TTEDGY.IFD	PRD.ITSDEG	D.AYVTPHMT	HSHWIKKDS	LSEAERAAQ
tr Q8DPQ2 Q8DPQ2_STRR6	TTSDGY.IFD	EHD.IISDEG	D.AYVTPHMG	HSHWIKKDS	LSDKEKVAAQ
tr Q9AG74 Q9AG74_STRPN	TTEDGY.IFD	PRD.ITSDEG	D.AYVTPHMT	HSHWIKKDS	LSEAERAAQ
tr Q9AHT9 Q9AHT9_STRPN	TTSDGY.IFD	EHD.IISDEG	D.AYVTPHMG	HSHWIKKDS	LSDKEKVAAQ
tr Q8DQ08 Q8DQ08_STRR6	TTEDGY.IFD	PRD.ITSDEG	D.AYVTPHMT	HSHWIKKDS	LSEAERAAQ



tr Q6WNQ5 Q6WNQ5_STRPN	GQKRVSFSFP	PELEKKLGIN	MLVKLITPDG	KVLEKVSQKV	FGEGVGNIAN
tr Q8CWR4 Q8CWR4_STRR6	AYAKEKGLTP	PSTDHQDSGN	.....TEA	KGAEAIYNRV	KAA.....KK
tr Q8DPQ2 Q8DPQ2_STRR6	AYTKEKGLP	PSPDADVKAN	.....PTG	DSAAAIYNRV	KGE.....KR
tr Q9AG74 Q9AG74_STRPN	AYAKEKGLTP	PSTDHQDSGN	.....TEA	KGAEAIYNRV	KAA.....KK
tr Q9AHT9 Q9AHT9_STRPN	AYTKEKGLP	PSPDADVKAN	.....PTG	DSAAAIYNRV	KGE.....KR
tr Q8DQ08 Q8DQ08_STRR6	AYAKEKGLTP	PSTDHQDSGN	.....TEA	KGAEAIYNRV	KAA.....KK

tr Q6WNQ5 Q6WNQ5_STRPN	FELDQPYLPG	QTFKYTIASK	DYPEVSYDGT	FTVPTSLAYK	MASQTIFYPF
tr Q8CWR4 Q8CWR4_STRR6	VPLDR..MP.	YNLQYTVVEVK	.....NGS	LIIP...HYD	HYHNIKFEWF
tr Q8DPQ2 Q8DPQ2_STRR6	IPLVR..LP.	YMVEHTVEVK	.....NGN	LIIP...HKD	HYHNIKFAWF
tr Q9AG74 Q9AG74_STRPN	VPLDR..MP.	YNLQYTVVEVK	.....NGS	LIIP...HYD	HYHNIKFEWF
tr Q9AHT9 Q9AHT9_STRPN	IPLVR..LP.	YMVEHTVEVK	.....NGN	LIIP...HKD	HYHNIKFAWF
tr Q8DQ08 Q8DQ08_STRR6	VPLDR..MP.	YNLQYTVVEVK	.....NGS	LIIP...HYD	HYHNIKFEWF

tr Q6WNQ5 Q6WNQ5_STRPN	HAGDTYLRVN	PQFAVPKGTD	ALVRVFDEFH	GNAYLENNYK	VGEIKLPIPK
tr Q8CWR4 Q8CWR4_STRR6	...DEGLYEA	PKGYSLEDLL	ATVKYYVE.H	PNRPHSDNG	FGNASDHVQR
tr Q8DPQ2 Q8DPQ2_STRR6	...DDHTYKA	PNGYTLEDLF	ATIKYYVE.H	PDERPHSNDG	WGNASEHVLG
tr Q9AG74 Q9AG74_STRPN	...DEGLYEA	PKGYSLEDLL	ATVKYYVE.H	PNRPHSDNG	FGNASDHVQR
tr Q9AHT9 Q9AHT9_STRPN	...DDHTYKA	PNGYTLEDLF	ATIKYYVE.H	PDERPHSNDG	WGNASEHVLG
tr Q8DQ08 Q8DQ08_STRR6	...DEGLYEA	PKGYSLEDLL	ATVKYYVE.H	PNRPHSDNG	FGNASDHVQR

tr Q6WNQ5 Q6WNQ5_STRPN	LNQGTTRTAG	NKIPVTFMAN	AYLDNQSTYI	VEVPILEKEN	QTD.....
tr Q8CWR4 Q8CWR4_STRR6	NKNGQADTNQ	TEKPNEEKPO	TEKPEEETPR	EEKPOSEKPE	S.....
tr Q8DPQ2 Q8DPQ2_STRR6	KKDHSEDPNK	NFKADEE...	.....	.....	.....
tr Q9AG74 Q9AG74_STRPN	NKNGQADTNQ	TEKPNEEKPO	TEKPEEETPR	EEKPOSEKPE	S.....
tr Q9AHT9 Q9AHT9_STRPN	KKDHSEDPNK	NFKADEE...	.....	.....	.....
tr Q8DQ08 Q8DQ08_STRR6	NKNGQADTNQ	TEKPNEEKPO	TEKPEEDKEH	DEVSEPTHE	SDEKENHVGL

tr Q6WNQ5 Q6WNQ5_STRPN	.....KP	SILPQFKRNK	AQENSKFDEK	VEEPTSEKV	EKEKLSETGN
tr Q8CWR4 Q8CWR4_STRR6	.P.....KP	...TEEPEEE	SPEES..PEE	SEEPQVETEK	VKEKLREA..
tr Q8DPQ2 Q8DPQ2_STRR6	.....P	.....VEET..PAE	PEVPQVETEK	VEAQLKEA..	
tr Q9AG74 Q9AG74_STRPN	.P.....KP	...TEEPEEE	SPEES..PEE	SEEPQVETEK	VKEKLREA..
tr Q9AHT9 Q9AHT9_STRPN	.....P	.....VEET..PAE	PEVPQVETEK	VEAQLKEA..	
tr Q8DQ08 Q8DQ08_STRR6	NPSADNLYKP	STDTEETEE	A. EDT..TDE	AEIPQVEHSV	INAKIAEA..

tr Q6WNQ5 Q6WNQ5_STRPN	STSNSTLEE	PTVDPVQEKV	AKFAESYGMK	LENVLFNMDG	TIELYLPSGE
tr Q8CWR4 Q8CWR4_STRR6	...EDLLGKI	Q..NPIIKSN	AKETLT.GLK	.NNLLFGTQD	NNTIMAEA..
tr Q8DPQ2 Q8DPQ2_STRR6	...EVLLAKV	T..DSSLKAN	ATETLA.GLR	.NNLTLQIMD	NNSIMAEA..
tr Q9AG74 Q9AG74_STRPN	...EDLLGKI	Q..NPIIKSN	AKETLT.GLK	.NNLLFGTQD	NNTIMAEA..
tr Q9AHT9 Q9AHT9_STRPN	...EVLLAKV	T..DSSLKAN	ATETLA.GLR	.NNLTLQIMD	NNSIMAEA..
tr Q8DQ08 Q8DQ08_STRR6	...EALLEKV	T..DSSIRQN	AVETLT.GLK	.SSLLLGTKD	NNTISAEV..

tr Q6WNQ5 Q6WNQ5_STRPN	VIKKNMADFT	GEAPQNGEN	KPSENGKVST	GTVENQPTEN	KPADSLPEAP
tr Q8CWR4 Q8CWR4_STRR6	..EKLLALLK	ESK.....	.....	.....	.....
tr Q8DPQ2 Q8DPQ2_STRR6	..EKLLALLK	GSPSSVSKE	KIN.....	.....	.....
tr Q9AG74 Q9AG74_STRPN	..EKLLALLK	ESK.....	.....	.....	.....
tr Q9AHT9 Q9AHT9_STRPN	..EKLLALLK	GSPSSVSKE	KIN.....	.....	.....
tr Q8DQ08 Q8DQ08_STRR6	..DSLLALLK	ESQPTPIQ..	.....	.....	.....

tr Q6WNQ5 Q6WNQ5_STRPN	NEKPVKPENS	TDNGMLNPEG	NVGSDPMLDP	ALEEAPAVDP	VQEKLEKFTA
tr Q8CWR4 Q8CWR4_STRR6	.....	.....	.....	.....	.....
tr Q8DPQ2 Q8DPQ2_STRR6	.....	.....	.....	.....	.....



tr Q9AG74 Q9AG74_STRPN	.....	.....	.....	.....	.....
tr Q9AHT9 Q9AHT9_STRPN	.....	.....	.....	.....	.....
tr Q8DQ08 Q8DQ08_STRR6	.....	.....	.....	.....	.....

tr Q6WNQ5 Q6WNQ5_STRPN	SYGLGLDSVI	FNMDGTIELR	LPSGEVIKKN	LSDLIA
tr Q8CWR4 Q8CWR4_STRR6	.....	.....	.....	.....
tr Q8DPQ2 Q8DPQ2_STRR6	.....	.....	.....	.....
tr Q9AG74 Q9AG74_STRPN	.....	.....	.....	.....
tr Q9AHT9 Q9AHT9_STRPN	.....	.....	.....	.....
tr Q8DQ08 Q8DQ08_STRR6	.....	.....	.....	.....

tr Q6WNQ7 Surface protein BVH-3 [bvh-3] [Streptococcus  
Q6WNQ7\_STRPN pneumoniae]

1039  
AA  
align

Score = 1134 bits (2933), Expect = 0.0  
Identities = 565/567 (99%), Positives = 565/567 (99%)

Query: 1 LTEEQIKAAQKHLEEVKTSNGLDSLSSHEQDYPSNAKEMKDLDKKIEEKIAGIMKQYGV 60  
LTEEQIKAAQKHLEEVKTSNGLDSLSSHEQDYF NAKEMEDLUKKIEEKIAGIMKQYGV  
Sbjct: 473 LTEEQIKAAQKHLEEVKTSNGLDSLSSHEQDYPGNAKEMKDLDKKIEEKIAGIMKQYGV 532

Query: 61 KRESIVVNKEKNALIIYPHGDHHHADPIDEHKPVGIGHSHSNYELFKPEEGVAKKEGNKVY 120  
KRESIVVNKEKNALIIYPHGDHHHADPIDEHKPVGIGHSHSNYELFKPEEGVAKKEGNKVY  
Sbjct: 533 KRESIVVNKEKNALIIYPHGDHHHADPIDEHKPVGIGHSHSNYELFKPEEGVAKKEGNKVY 592

Query: 121 TGEELTNVVNLLKNSTFNNQNFTLANGQKRVSFSFPPELEKKLGINMLVKLITPDGKVLE 180  
TGEELTNVVNLLKNSTFNNQNFTLANGQKRVSFSFPPELEKKLGINMLVKLITPDGKVLE  
Sbjct: 593 TGEELTNVVNLLKNSTFNNQNFTLANGQKRVSFSFPPELEKKLGINMLVKLITPDGKVLE 652

Query: 181 KVSGKVFGEVGNIANFELDQPYLPGQTFKYTIASKDYPEVSYDGTFTVPTSLAYKMASQ 240  
KVSGKVFGEVGNIANFELDQPYLPGQTFKYTIASKDYPEVSYDGTFTVPTSLAYKMASQ  
Sbjct: 653 KVSGKVFGEVGNIANFELDQPYLPGQTFKYTIASKDYPEVSYDGTFTVPTSLAYKMASQ 712

Query: 241 TIFYPFHAGD TYLRVNPQFAVPKGT DALVRVFDEFHGNAYLENNYKVGEIKLPIPKLNQG 300  
TIFYPFHAGD TYLRVNPQFAVPKGT DALVRVFDEFHGNAYLENNYKVGEIKLPIPKLNQG  
Sbjct: 713 TIFYPFHAGD TYLRVNPQFAVPKGT DALVRVFDEFHGNAYLENNYKVGEIKLPIPKLNQG 772

Query: 301 TTRTAGNKIPVTFMANAYLDNQSTYIVEVPILEKENQTDKPSILPQFKRKAQENLKLDE 360  
TTRTAGNKIPVTFMANAYLDNQSTYIVEVPILEKENQTDKPSILPQFKRKAQEN KLDE  
Sbjct: 773 TTRTAGNKIPVTFMANAYLDNQSTYIVEVPILEKENQTDKPSILPQFKRKAQENSKLDE 832

Query: 361 KVEEPTSEKVEKEKLSETGNSTSNSTLEEVPTVDPVQEKVAKFAESYGMKLENVLFNMD 420  
KVEEPTSEKVEKEKLSETGNSTSNSTLEEVPTVDPVQEKVAKFAESYGMKLENVLFNMD  
Sbjct: 833 KVEEPTSEKVEKEKLSETGNSTSNSTLEEVPTVDPVQEKVAKFAESYGMKLENVLFNMD 892

Query: 421 GTIELYLPSEGEVIKKNMADFTGEAPQNGENKPSSENGKVSTGTVENQPTENKPADSLPEA 480  
GTIELYLPSEGEVIKKNMADFTGEAPQNGENKPSSENGKVSTGTVENQPTENKPADSLPEA  
Sbjct: 893 GTIELYLPSEGEVIKKNMADFTGEAPQNGENKPSSENGKVSTGTVENQPTENKPADSLPEA 952

Query: 481 PNEKPVKPENSTDNGMLNPEGNVGSDPMLDPALEEAPAVDPVQEKLEKFTASYGLGLDSV 540  
PNEKPVKPENSTDNGMLNPEGNVGSDPMLDPALEEAPAVDPVQEKLEKFTASYGLGLDSV  
Sbjct: 953 PNEKPVKPENSTDNGMLNPEGNVGSDPMLDPALEEAPAVDPVQEKLEKFTASYGLGLDSV 1012

Query: 541 IFNMDGTIELRLPSGEVIKKNLSDLIA 567  
IFNMDGTIELRLPSGEVIKKNLSDLIA  
Sbjct: 1013 IFNMDGTIELRLPSGEVIKKNLSDLIA 1039

Q6WNQ5 Surface protein BVH-3 (Fragment) [bvh-3] [Streptococcus 1019  
Q6WNQ5\_STRPN pneumoniae] AA  
align

Score = 1504 bits (3893), Expect = 0.0  
Identities = 743/779 (95%), Positives = 743/779 (95%)

*phTE*  
*BVH-3*

Query: 1 AYALNQHRSQENKDNRRVSYVDGSQSSQKSENLTDPQVSQKEGIQAEQIVIKITDQGYVT 60  
AYALNQHRSQENKDNRRVSYVDGSQSSQKSENLTDPQVSQKEGIQAEQIVIKITDQGYVT  
Sbjct: 2 AYALNQHRSQENKDNRRVSYVDGSQSSQKSENLTDPQVSQKEGIQAEQIVIKITDQGYVT 80

Query: 61 SHGDHYHYINGKVPYDALFSEELLMKDPNYQLKDADIVNEVKGGYIIKVDGKYYVYLKDA 120  
SHGDHYHYINGKVPYDALFSEELLMKDPNYQLKDADIVNEVKGGYIIKVDGKYYVYLKDA  
Sbjct: 62 SHGDHYHYINGKVPYDALFSEELLMKDPNYQLKDADIVNEVKGGYIIKVDGKYYVYLKDA 140

Query: 121 AHADNVRTKDEINRQKQEHVKDNEKVNNSVAVARSQGRYTTNDGYVFNPAIIEDTGNAY 180  
AHADNVRTKDEINRQKQEHVKDNEKVNNSVAVARSQGRYTTNDGYVFNPAIIEDTGNAY  
Sbjct: 122 AHADNVRTKDEINRQKQEHVKDNEKVNNSVAVARSQGRYTTNDGYVFNPAIIEDTGNAY 200

Query: 181 IVPHGHHYHYIPXXXXXXXXXXXXXXXXXXXXNMQPSQLSYSSTASDNNTQSVAKGSTSKP 240  
IVPH GHYHYIP NMQPSQLSYSSTASDNNTQSVAKGSTSKP  
Sbjct: 182 IVPHRGHYHYIPKSDLSASELAAKAHLAGKNMQPSQLSYSSTASDNNTQSVAKGSTSKP 260

Query: 241 ANKSENLSLLKELYDSPAQRYSSES DGLVFDPAKIIISRTPNGVAIPHGDHYHFIPYSKL 300  
ANKSENLSLLKELYDSPAQRYSSES DGLVFDPAKIIISRTPNGVAIPHGDHYHFIPYSKL  
Sbjct: 242 ANKSENLSLLKELYDSPAQRYSSES DGLVFDPAKIIISRTPNGVAIPHGDHYHFIPYSKL 320

Query: 301 SALEEKIARMVPISGTGSTVSTNAKPNEVVXXXXXXXXXXXXXXXXXXXXKELSSASDGYIFNP 360  
SALEEKIARMVPISGTGSTVSTNAKPNEVV KELSSASDGYIFNP  
Sbjct: 302 SALEEKIARMVPISGTGSTVSTNAKPNEVVSSLGSLSSNPSSLTTSKELSSASDGYIFNP 380

Query: 361 KDIVEETATAYIVRHGDHFHYIPKSNQIGQPTLPNNSLATPSPSLPINPGTSHEKHEEDG 420  
KDIVEETATAYIVRHGDHFHYIPKSNQIGQPTLPNNSLATPSPSLPINPGTSHEKHEEDG  
Sbjct: 362 KDIVEETATAYIVRHGDHFHYIPKSNQIGQPTLPNNSLATPSPSLPINPGTSHEKHEEDG 440

Query: 421 YGFDANRIIAEDES GFVMSHGDHNHYFFKKDLTEEQIKAAQKHLEEVKTS HNGLDLSLSSH 480  
YGFDANRIIAEDES GFVMSHGDHNHYFFKKDLTEEQIKAAQKHLEEVKTS HNGLDLSLSSH  
Sbjct: 422 YGFDANRIIAEDES GFVMSHGDHNHYFFKKDLTEEQIKAAQKHLEEVKTS HNGLDLSLSSH 500

Query: 481 EQDYPSNAKEMKDLDDKKIEEKIAGIMKQYGVKRESIVVNKEKNALIIYPHGDHHDADPIDE 540  
EQDYPSNAKEMKDLDDKKIEEKIAGIMKQYGVKRESIVVNKEKNALIIYPHGDHHDADPIDE  
Sbjct: 482 EQDYPSNAKEMKDLDDKKIEEKIAGIMKQYGVKRESIVVNKEKNALIIYPHGDHHDADPIDE 560

Query: 541 HKPVGIGHSHSNYELFKPEEGVAKKEGKNKYVTGEELTNVVNLLKNSTFNNQNFTLANGQK 600  
HKPVGIGHSHSNYELFKPEEGVAKKEGKNKYVTGEELTNVVNLLKNSTFNNQNFTLANGQK  
Sbjct: 542 HKPVGIGHSHSNYELFKPEEGVAKKEGKNKYVTGEELTNVVNLLKNSTFNNQNFTLANGQK 620

Query: 601 RVSFSPPELEKKLGINMLVKLITPDGKVLEKVS GKVFGEGVGNIANFELDQPYLPQGTF 660  
RVSFSPPELEKKLGINMLVKLITPDGKVLEKVS GKVFGEGVGNIANFELDQPYLPQGTF  
Sbjct: 602 RVSFSPPELEKKLGINMLVKLITPDGKVLEKVS GKVFGEGVGNIANFELDQPYLPQGTF 680

Query: 661 KYTIASKDYPEVSYDGTFTVPTSLAYKMASQTI FYPFHAGDTYLRVNPQFAVPKGTDALV 720  
KYTIASKDYPEVSYDGTFTVPTSLAYKMASQTI FYPFHAGDTYLRVNPQFAVPKGTDALV  
Sbjct: 662 KYTIASKDYPEVSYDGTFTVPTSLAYKMASQTI FYPFHAGDTYLRVNPQFAVPKGTDALV 740

Query: 721 RVFDEFHGNAYLENNYKVGEIKLPIPKLNQGTTRTAGNKIPVTFMANAYLDNQSTYIVE 779  
RVFDEFHGNAYLENNYKVGEIKLPIPKLNQGTTRTAGNKIPVTFMANAYLDNQSTYIVE  
Sbjct: 722 RVFDEFHGNAYLENNYKVGEIKLPIPKLNQGTTRTAGNKIPVTFMANAYLDNQSTYIVE 800

tr Q6WNQ7 Surface protein BVH-3 [bvh-3] [Streptococcus  
Q6WNQ7\_STRPN pneumoniae]

1039  
AA  
align

Score = 475 bits (1222), Expect = e-133  
Identities = 239/240 (99%), Positives = 239/240 (99%)

phTE → Query: 1 EVPILEKENQTDKPSILPQFKRNKAQENLKLDEKVVEPKTSEKVEKEKLSETGNSTSNST 60  
BVH 3 → Sbjct: 800 EVPILEKENQTDKPSILPQFKRNKAQEN KLDEKVVEPKTSEKVEKEKLSETGNSTSNST 859  
Query: 61 LEEVPTVDPVQEKVAKFAESYGMKLENVLFNMDGTIELYLPSGEVIKKNMADFTGEAPQG 120  
Sbjct: 860 LEEVPTVDPVQEKVAKFAESYGMKLENVLFNMDGTIELYLPSGEVIKKNMADFTGEAPQG 919  
Query: 121 NGENKPSSENGKVSTGTVENQPTENKPADSLPEAPNEKPVKPENSTDNGMLNPEGNGVSDP 180  
Sbjct: 920 NGENKPSSENGKVSTGTVENQPTENKPADSLPEAPNEKPVKPENSTDNGMLNPEGNGVSDP 979  
Query: 181 MLDPALIEEAPAVDPVQEKLEKFTASYGLGLDSVIFNMDGTIELRLPSGEVIKKNLSDLIA 240  
Sbjct: 980 MLDPALIEEAPAVDPVQEKLEKFTASYGLGLDSVIFNMDGTIELRLPSGEVIKKNLSDLIA 1039

tr Q6WNQ7 Surface protein BVH-3 [bvh-3] [Streptococcus  
Q6WNQ7\_STRPN pneumoniae]

1039  
AA  
align

Score = 1059 bits (2738), Expect = 0.0  
Identities = 527/528 (99%), Positives = 527/528 (99%)

*PH+E*  
*BV#3*

Query: 1 MKDLDDKKIEEKIAGIMKQYGVKRESIVVNKEKNAIIPHGDDHHADPIDEHKPVGIGHSH 60  
MKDLDDKKIEEKIAGIMKQYGVKRESIVVNKEKNAIIPHGDDHHADPIDEHKPVGIGHSH  
Sbjct: 512 MKDLDDKKIEEKIAGIMKQYGVKRESIVVNKEKNAIIPHGDDHHADPIDEHKPVGIGHSH 571

Query: 61 SNYELFKPEEGVAKKEGKNKVYTGEELTNVVNLLKNSTFNNQNFTLANGQKRVSFSFPPEL 120  
SNYELFKPEEGVAKKEGKNKVYTGEELTNVVNLLKNSTFNNQNFTLANGQKRVSFSFPPEL  
Sbjct: 572 SNYELFKPEEGVAKKEGKNKVYTGEELTNVVNLLKNSTFNNQNFTLANGQKRVSFSFPPEL 631

Query: 121 EKKLGINMLVVKLITPDGKVLKVSQKVFGEVGNIANFELDQPYLPGQTFKYTIASKDYP 180  
EKKLGINMLVVKLITPDGKVLKVSQKVFGEVGNIANFELDQPYLPGQTFKYTIASKDYP  
Sbjct: 632 EKKLGINMLVVKLITPDGKVLKVSQKVFGEVGNIANFELDQPYLPGQTFKYTIASKDYP 691

Query: 181 EVSYDGTFTVPTSLAYKMASQTIFYPFHAGDTYLRVNPQFAVPKGTDALVRVFEDEFHGNA 240  
EVSYDGTFTVPTSLAYKMASQTIFYPFHAGDTYLRVNPQFAVPKGTDALVRVFEDEFHGNA  
Sbjct: 692 EVSYDGTFTVPTSLAYKMASQTIFYPFHAGDTYLRVNPQFAVPKGTDALVRVFEDEFHGNA 751

Query: 241 YLENNYKVGEIKLPIPKLNQGTTRTAGNKIPVTFMANAYLDNQSTYIVEVPILEKENQTD 300  
YLENNYKVGEIKLPIPKLNQGTTRTAGNKIPVTFMANAYLDNQSTYIVEVPILEKENQTD  
Sbjct: 752 YLENNYKVGEIKLPIPKLNQGTTRTAGNKIPVTFMANAYLDNQSTYIVEVPILEKENQTD 811

Query: 301 KPSILPQFKRNKAQENLKLDEKVEEPTSEKVEKEKLSETGNSTSNSTLEEVPVDPVQE 360  
KPSILPQFKRNKAQEN KLDEKVEEPTSEKVEKEKLSETGNSTSNSTLEEVPVDPVQE  
Sbjct: 812 KPSILPQFKRNKAQENSKLDEKVEEPTSEKVEKEKLSETGNSTSNSTLEEVPVDPVQE 871

Query: 361 KVAKFAESYGMKLENVLFNMDGTIELYLPSGEVIKKNMADFTGEAPQGNGENKPSSENGKV 420  
KVAKFAESYGMKLENVLFNMDGTIELYLPSGEVIKKNMADFTGEAPQGNGENKPSSENGKV  
Sbjct: 872 KVAKFAESYGMKLENVLFNMDGTIELYLPSGEVIKKNMADFTGEAPQGNGENKPSSENGKV 931

Query: 421 STGTVENQPTENKPADSLPEAPNEKPVKPENSTDNGMLNPEGNGVSDPMLDPALEEAPAV 480  
STGTVENQPTENKPADSLPEAPNEKPVKPENSTDNGMLNPEGNGVSDPMLDPALEEAPAV  
Sbjct: 932 STGTVENQPTENKPADSLPEAPNEKPVKPENSTDNGMLNPEGNGVSDPMLDPALEEAPAV 991

Query: 481 DPVQEKLEKFTASYGLGLDSVIFNMDGTIELRLPSGEVIKKNLSDLIA 528  
DPVQEKLEKFTASYGLGLDSVIFNMDGTIELRLPSGEVIKKNLSDLIA  
Sbjct: 992 DPVQEKLEKFTASYGLGLDSVIFNMDGTIELRLPSGEVIKKNLSDLIA 1039

[ExPASy Home page](#)[Site Map](#)[Search ExPASy](#)[Contact us](#)[Swiss-Prot](#)Search for 

# UniProtKB/TrEMBL

## entry Q9ANY1

[Printer-friendly view](#)[Request update](#)[Q1](#)

[\[Entry info\]](#)
[\[Name and origin\]](#)
[\[References\]](#)
[\[Comments\]](#)
[\[Cross-references\]](#)
[\[Keywords\]](#)  
[\[Features\]](#)
[\[Sequence\]](#)
[\[Tools\]](#)

*Note: most headings are clickable, even if they don't appear as links. They link to the user manual or other documents.*

### Entry information

Entry name	<b>Q9ANY1_STRPN</b>
Primary accession number	<b>Q9ANY1</b>
Secondary accession number	<b>Q7D4B6</b>
Entered in TrEMBL in	Release 17, June 2001
Sequence was last modified in	Release 17, June 2001
Annotations were last modified in	Release 30, May 2005
<b>Name and origin of the protein</b>	
Protein name	<b>Pneumococcal histidine triad protein E [Precursor]</b>
Synonym	<b>Hypothetical protein SP1004</b>
Gene name	<b>Name: phtE</b>
	OrderedLocusNames: SP1004
From	Streptococcus pneumoniae [TaxID: 1313]
Taxonomy	Bacteria; Firmicutes; Lactobacillales; Streptococcaceae; Streptococcus.

### References

- [1] NUCLEOTIDE SEQUENCE.  
 DOI=10.1128/IAI.69.2.949-958.2001; PubMed=11159990 [NCBI, ExPASy, EBI, Israel, Japan]  
 Adamou J.E., Heinrichs J.H., Erwin A.L., Walsh W., Gayle T., Dormitzer M., Dagan R., Brewah Y.A., Barren P., Lathigra R., Langermann S., Koenig S., Johnson S.;  
 "Identification and characterization of a novel family of pneumococcal proteins (the Pht family) that are protective against sepsis.";  
 Infect. Immun. 69:949-958(2001).
- [2] NUCLEOTIDE SEQUENCE.  
**STRAIN=ATCC BAA-334 / TIGR4;**  
 DOI=10.1126/science.1061217; PubMed=11463916 [NCBI, ExPASy, EBI, Israel, Japan]  
 Tettelin H., Nelson K.E., Paulsen I.T., Eisen J.A., Read T.D., Peterson S.N., Heidelberg J.F., DeBoy R.T., Haft D.H., Dodson R.J., Durkin A.S., Gwinn M.L., Kolonay J.F., Nelson W.C., Peterson J.D., Umayam L.A., White O., Salzberg S.L., Lewis M.R., Fraser C.M.;  
 "Complete genome sequence of a virulent isolate of Streptococcus pneumoniae.";  
 Science 293:498-506(2001).

### Comments

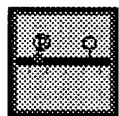
None

### Cross-references

AF318956; AAK06761.1; -;  
 Genomic\_DNA.

[EMBL / GenBank / DDBJ]  
 [CoDingSequence]

EMBL AE007403; AAK75121.1; -; [EMBL / GenBank / DDBJ]  
 Genomic\_DNA. [CoDingSequence]  
 PIR H95115; H95115.  
 TIGR SP1004; -.  
 InterPro IPR006270; Strep\_his\_triad.  
 Graphical view of domain structure.  
 Pfam PF04270; Strep\_his\_triad; 5.  
 Pfam graphical view of domain structure.  
 TIGRFAMs TIGR01363; strep\_his\_triad; 3.  
 ProDom [Domain structure / List of seq. sharing at least 1 domain]  
 HOGENOM [Family / Alignment / Tree]  
 ProtoMap Q9ANY1.  
 PRESAGE Q9ANY1.  
 ModBase Q9ANY1.  
 SWISS-2DPAGE Get region on 2D PAGE.  
 UniRef View cluster of proteins with at least 50% / 90% identity.

**Keywords****Complete proteome; Hypothetical protein; Signal.****Features**

Feature table viewer

Key	From	To	Length	Description
SIGNAL	1	29	29	Potential.

**Sequence information**

Length: 1039 Molecular weight: 114631 CRC64: 81A563FC806625C4 [This is a checksum on the  
 AA Da sequence]

<u>10</u>	<u>20</u>	<u>30</u>	<u>40</u>	<u>50</u>	<u>60</u>
MKFSKKYIAA	GSAVIVLSL	CAYALNQHR	S QENKDNRRVS	YVDGSQSSQK	SENLTDPQVS
<u>70</u>	<u>80</u>	<u>90</u>	<u>100</u>	<u>110</u>	<u>120</u>
QKEGIQAEQI	VIKITDQGYV	TSHGDHYHY	Y NGKVPYDALF	SEELLMKDPN	YQLKDADIVN
<u>130</u>	<u>140</u>	<u>150</u>	<u>160</u>	<u>170</u>	<u>180</u>
EVKGGYIIKV	DGKYVYLKD	AAHADNVRTK	DEINRQKQEH	VKDNEKVNSN	VAVARSQGRY
<u>190</u>	<u>200</u>	<u>210</u>	<u>220</u>	<u>230</u>	<u>240</u>
TTNDGYVFNP	ADIIEDTGNA	YIVPHGGHYH	Y IPKSDLSAS	ELAAAKAHLA	GKNMQPSQLS
<u>250</u>	<u>260</u>	<u>270</u>	<u>280</u>	<u>290</u>	<u>300</u>
YSSTASDNNT	QSVAKGSTSK	PANKSENLOS	LLKELYDSPA	AQRYSESDGL	VFDPAKIISR
<u>310</u>	<u>320</u>	<u>330</u>	<u>340</u>	<u>350</u>	<u>360</u>
TPNGVAIPHG	DHYHFIPYSK	LSALEEKIAR	MVPISGTGST	VSTNAKPNEV	VSSLGSLSSN
<u>370</u>	<u>380</u>	<u>390</u>	<u>400</u>	<u>410</u>	<u>420</u>
PSSLTTSKEL	SSASDGYIFN	PKDIVEETAT	AYIVRHGDHF	HYIPKSNQIG	QPTLPNNSLA
<u>430</u>	<u>440</u>	<u>450</u>	<u>460</u>	<u>470</u>	<u>480</u>

```

TPSPSLPINP GTSHEKHEED GYGFDANRII AEDESGFVMS HGDHNHYFFK KDLTEEQIKA
      490      500      510      520      530      540
AQKHLEEVKT SHNGLDSLSS HEQDYPSTNAK EMKDLDKKIE EKIAGIMKQY GVKRESIVVN
      550      560      570      580      590      600
KEKNAIIYPH GDHHHADPID EHKPVGIGHS HSNYELFKPE EGVAKKEGNK VYTGEELTNV
      610      620      630      640      650      660
VNLLKNSTFN NQNFTLANGQ KRVSFSFPPE LEKKLGINML VKLITPDGKV LEKVSQKVFQ
      670      680      690      700      710      720
EGVGNIANFE LDQPYLPGQT FKYTIASKDY PEVSYDGTFT VPTSLAYKMA SQTIFYPFHA
      730      740      750      760      770      780
GDTYLRVNPQ FAVPKGTDAL VRFDEFHGN AYLENNYKVG EIKLPIPKLN QGTTRTAGNK
      790      800      810      820      830      840
IPVTFMANAY LDNQSTYIVE VPILEKENQT DKPSILPQFK RNKAQENLKL DEKVEEPKTS
      850      860      870      880      890      900
EKVEKEKLSE TGNSTSNSTL EEVPTVDPVQ EKVAKFAESY GMKLENVLFN MDGTIELYLP
      910      920      930      940      950      960
SGEVIKKNMA DFTGEAPQGN GENKPSENGK VSTGTVENQP TENKPADSLP EAPNEKPVKP
      970      980      990     1000     1010     1020
ENSTDNGMLN PEGNVGSDPM LDPALEEAPA VDPVQEKLEK FTASYGLGLD SVIFNMDGTI
      1030
ELRLPSGEVI KKNLSDLIA

```

Q9ANY1 in FASTA  
format

*View entry in original UniProtKB/TrEMBL format*

*View entry in raw text format (no links)*

*Request for annotation of this UniProtKB/TrEMBL entry*

**BLAST** BLAST submission on  
ExPASy/SIB  
or at NCBI (USA)



Sequence analysis tools: ProtParam, ProtScale,  
Compute pI/Mw, PeptideMass, PeptideCutter,  
Dotlet (Java)



ScanProsite, MotifScan



Submit a homology modeling request to SWISS-  
MODEL



NPSA Sequence analysis  
tools



ExPASy Home page

Site Map

Search ExPASy

Contact us

Swiss-Prot

Hosted by  NHRI Taiwan Mirror sites: Australia Brazil Canada Korea Switzerland USA